BIOINFORMATICS

Vol. 15 no. 5 1999
Pages 391-412

Automated genome sequence analysis and annotation

Miguel A. Andrade[1], Nigel P. Brown[1], Christophe Leroy[1], Sebastian Hoersch[1], Antoine de Daruvar[1], Christian Reich[1], Angelo Franchini[1], Javier Tamames[2], Alfonso Valencia[2], Christos Ouzounis[1] and Chris Sander[1,*]

[1]European Bioinformatics Institute (EBI), Wellcome Trust Genome Campus, Cambridge CB10 1SD, UK and [2]Protein Design Group, CNB-CSIC, Campus U. Autónoma, Cantoblanco, Madrid M-28049, Spain

Received on November 26, 1998; revised on January 15, 1999; accepted on January 18, 1999

Abstract

Motivation: *Large-scale genome projects generate a rapidly increasing number of sequences, most of them biochemically uncharacterized. Research in bioinformatics contributes to the development of methods for the computational characterization of these sequences. However, the installation and application of these methods require experience and are time consuming.*

Results: *We present here an automatic system for preliminary functional annotation of protein sequences that has been applied to the analysis of sets of sequences from complete genomes, both to refine overall performance and to make new discoveries comparable to those made by human experts. The GeneQuiz system includes a Web-based browser that allows examination of the evidence leading to an automatic annotation and offers additional information, views of the results, and links to biological databases that complement the automatic analysis. System structure and operating principles concerning the use of multiple sequence databases, underlying sequence analysis tools, lexical analyses of database annotations and decision criteria for functional assignments are detailed. The system makes automatic quality assessments of results based on prior experience with the underlying sequence analysis tools; overall error rates in functional assignment are estimated at 2.5–5% for cases annotated with highest reliability ('clear' cases). Sources of over-interpretation of results are discussed with proposals for improvement. A conservative definition for reporting 'new findings' that takes account of database maturity is presented along with examples of possible kinds of discoveries (new function, family and superfamily) made by the system. System performance in relation to sequence database coverage, database dynamics and database search methods is analysed, demonstrating the inherent advantages of an integrated automatic approach using multiple databases and search methods applied in an objective and repeatable manner.*

Availability: *The GeneQuiz system is publicly available for analysis of protein sequences through a Web server at http://www.sander.ebi.ac.uk/gqsrv/submit*

Contact: *sander@mpi.com*

Supplementary information: *http://www.sander.ebi.ac.uk/genequiz/*

Introduction

Functional analyses of protein sequences can now be performed on a computer using a variety of software tools that allow the user to exploit the biochemical knowledge accumulated in sequence databases. For example, the correlation of sequence similarity with similarity of function provides a basis for transferring functional knowledge from a biochemically characterized protein to a homologous, but otherwise uncharacterized one. Given a protein sequence, analysis of the conservation patterns in the corresponding protein family can allow the association of regions of the sequence or of individual residues with structural or functional motifs and may even allow the construction of a three-dimensional (3D) model by homology to a known structure in the family. Such theoretically obtained functional and structural insights may be used to direct the comparatively much more lengthy, difficult and expensive experimentation on the real protein.

Although these methods are available to the researcher, their application can be cumbersome for various reasons. First, computer programs may be difficult to install and maintain. Some of them require the combined installation of huge nucleotide and protein databases that currently contain hundreds of thousands of sequences requiring gigabytes of disk storage space. The installation and maintenance of such

*Send correspondence to C.Sander, Whitehead Institute, MIT Center for Genome Research, Cambridge, MA 02139, USA programs and/or databases require suitably powerful computer hardware as well as special skills, so that the effort may be disproportionate for an experimental group working on a small number of proteins. Fortunately, for small requirements, some of these tools are available for interactive (Web server) or semi-interactive (Web or mail server) use over the Internet. However, the user will be constrained by the variety of software available in this manner, as well as by the choice of databases or even program parameters provided by any service, and by the limiting turnaround time of the remote service or the speed of Internet access.

Even if access to appropriate software and databases is available, a second major difficulty is the need for specialist skills in using these programs effectively, both through the appropriate choice of controlling parameter settings and in evaluating the significance of the results. This expert knowledge can only be acquired through repeated use of the tools, often comparing and combining results from several methods. Again, a researcher interested only in a small number of proteins may not have this experience.

If a group is interested in analysing a great number of uncharacterized sequences, as from the large-scale sequencing projects, then installation of the programs and databases and investment in the necessary expertise are worthwhile, indeed essential. However, a third problem arises, namely the application of the methods and evaluation of the results for a large number of sequences require a considerable amount of computer and human expert time, as well as tight quality control to ensure a uniformity of application and interpretation. Moreover, methods and databases improve over time and frequent re-analysis may bring new results.

A partial solution to these three problems, (i) flexible installation and maintenance of a set of methods and databases, (ii) need for expertise in the use and evaluation of the methods and (iii) fast and uniform analysis of the results, was addressed with the development of the first GeneQuiz system (Scharf *et al*., 1994; Casari *et al*., 1996).

GeneQuiz is a semi-automated protein sequence analysis system, the principal purpose of which is to infer a specific and reliable functional assignment together with a broad cellular role for a query protein by analysis of annotations from sequence database matches. The system also applies a selected suite of analysis tools to the query sequence, integrating the results into a coherent display to complement the functional assignments.

The GeneQuiz system is able to process large numbers of sequences quickly and repeatably in a consistent manner, and makes use of regularly updated combined sequence databases. Thus, the system can be used for occasional analyses of a few query protein sequences, or it can be systematically applied to the large numbers of open reading frames (ORFs) identified in a genome sequencing project.

A high degree of automation is required to cope with the analysis of the huge number of sequences generated by genome sequencing projects, and to ensure consistent and reproducible results, freeing the expert user to verify and refine these analyses and to follow up new discoveries. Another advantage of a high-throughput system is that, because the analysis of a genome is not yet a stable problem, it must be periodically repeated to utilize the constantly increasing information held in biological databases.

In summary, the GeneQuiz system may be viewed as a protein sequence analysis workbench with the primary goal of automatic functional inference, and a secondary goal of presentation of supporting information abstracted from the different sequence analyses.

GeneQuiz has been used in analyses of individual proteins and of complete genomes: *Haemophilus influenzae* (Casari *et al*., 1995b), *Mycoplasma genitalium* Rd (Ouzounis *et al*., 1996b), *Methanococcus jannaschii* (Andrade *et al*., 1997) and others (see the Web site at http://www.sander.ebi.ac.uk/genequiz/). The extensive experience thus obtained with GeneQuiz has been fed back into improvements to the system and the addition of new features.

The latest improvements have mainly focused on those parts of the system concerned with reasoning about protein function and on the user interface. The reasoning module now includes a lexical analysis of the description fields of sequences homologous to a given query sequence, that better discriminates sequence annotations with functional content from those without. A completely new browsing module displays the full analysis of a query sequence as a one-page hyperlinked report, offering, in particular, graphical views of collated homologous fragments, sequence and structure motifs, and predicted structural features aligned with the query.

Finally, sequence analysis using the GeneQuiz system has now been made available to the community through a World Wide Web server. Protein sequences can be submitted for analysis and the complete results may be browsed on the Web.

This paper presents these developments in the context of a full description of the complete system. The performance of GeneQuiz is then evaluated with some examples, and the problems of this type of approach to function inference are discussed with suggestions for possible solutions. Finally, the implications of automatic systems such as GeneQuiz in the field of genome analysis and protein function annotation are discussed, again suggesting future directions for development.

The GeneQuiz system

The GeneQuiz system takes as input a protein sequence and produces as output a specific functional annotation and general functional class for this sequence. The user can browse the results of the analysis and additional information that can

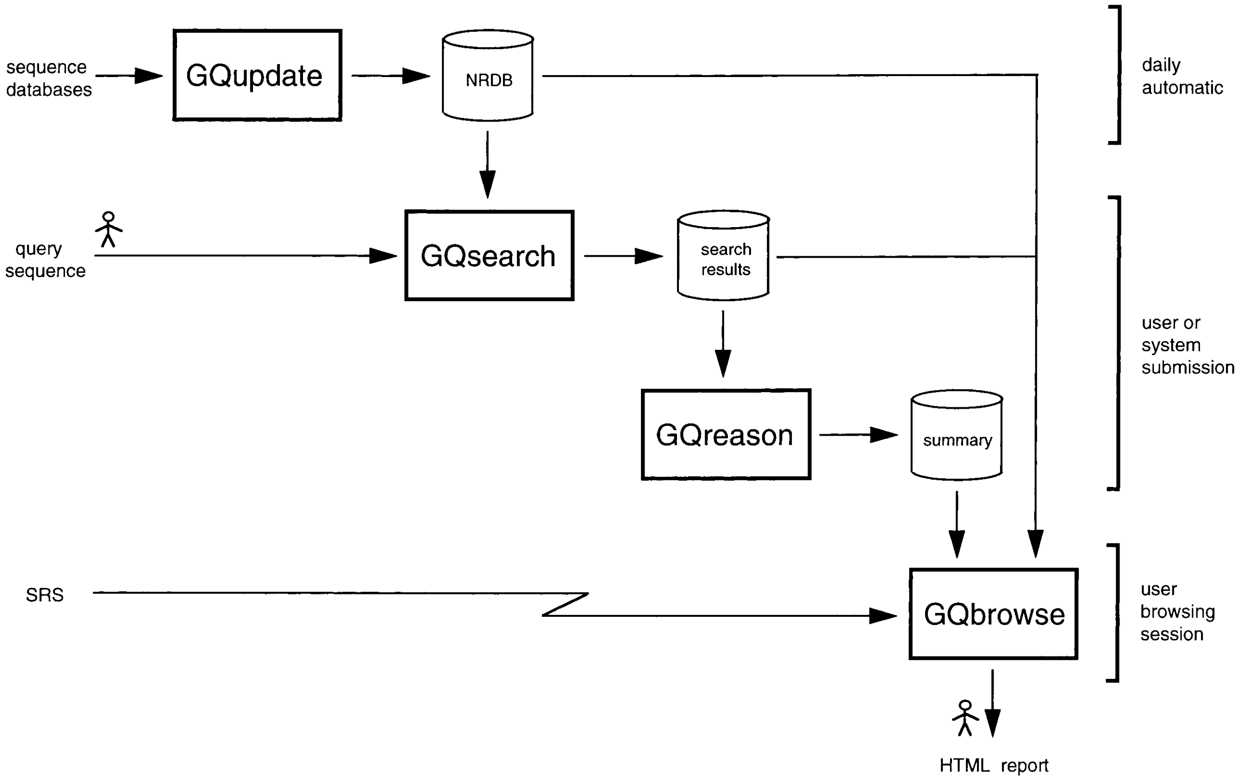


Fig. 1. GeneQuiz modules and control flow. Boxes depict the four GeneQuiz modules, while cylinders represent data storage. Inputs at l eft are raw sequences from the public databases, the query sequence for a single run, and links to external database annotations via SRS. Output in Web-browsable form is at the bottom. Stick figures indicate user interaction. An indication of the frequency of operation of the various subsystems is given at right. NRDB, non-redundant database of protein sequences; SRS, Sequence Retrieval System (Etzold *et al*., 1996); HTML, Hypertext Markup Language.

be used to confirm the automatic annotation or make new deductions.

GeneQuiz is composed of four modules: GQupdate, GQsearch, GQreason and GQbrowse, which are shown in Figure 1 and described in subsequent sections. The GQupdate module is responsible for maintaining integrated, up-to-date, non-redundant protein and nucleotide sequence databases derived from a compendium of public databases, as well as databases of protein structures and motifs. Extensive use of these databases is made by the other modules.

A single GeneQuiz run is triggered by entry of a query protein sequence into the system, either by a user submission, or as one of a batch of sequences, perhaps representing the protein set of a full genome. The GQsearch module applies a variety of sequence analysis tools to the query sequence, parsing, and storing the results in a common format for subsequent processing stages. In particular, the query is screened against the non-redundant sequence databases using several standard database search programs and a multiple alignment is constructed.

GQreason uses these results together with the original database annotations in the form of keywords and sequence descriptions to assign, where possible, a specific function to the query by transfer of function from a homologue, a general functional class for the homologues grouped as a family, and a reliability estimate for the procedure.

Finally, GQbrowse allows the user to examine through a Web browser the derived conclusions, the evidence that led to the functional annotation, graphical displays of alignments and 3D models, supplementary information from other tools, and the original sequence analysis output, all with links to the external public sequence and motif databases via Sequence Retrieval System (SRS) (Etzold *et al*., 1996).

Module: GQupdate

In order to extract the utmost information from the various biological databases, it is essential to have a complete and up-to-date collection of well-annotated database entries. The function of the GQupdate module is, therefore, to gather all

Table 1. Databases managed by GQupdate (December 1997). Database sizes are given as the number of sequences

Content	Database	Release	Date	Size	NRDB size
Protein sequences	SWISS-PROT[1]	34.0	10/96	59,021	276,481
	SWISSnew[1]	–	–	27,112	
	PIR[2]	53.0	8/97	95,051	
	WormPep[3]	12	11/97	12,178	
Inferred translations of nucleotide sequences	TREMBL[1]	EMBL	8/97	265,695	
	TREMBLnew[1]	–	–	42,017	
	GenPept[4]	Genbank	10/97	262,153	
	GenPeptnew[4]	–	–	29,732	
Nucleotide sequences	EMBL[5]	52	10/97	1,787,004	1,331,154
	EMBLnew[5]	–	–	187,343	
	Genbank[4]	103.0	10/97	1,093,244	
	Genbanknew[4]	–	–	61,783	
Protein motifs	Blocks[6]	9.3	3/97	3,417	–
	PROSITE[7]	14.0	11/97	1,167	
Protein structures	PDB[8]	–	12/97	6,664	

NRDB, non-redundant sequence database.
[1]Bairoch and Apweiler (1997); [2]George *et al.* (1997); [3]WormPep databank; [4]Benson *et al.* (1997); [5]Stoesser *et al.* (1997); [6]Henikoff *et al.* (1997); [7]Bairoch *et al.* (1997); [8]Abola *et al.* (1987).

new sequences or other entries as they appear in the public databases and to merge these into the repository used by the GeneQuiz engine.

GQupdate operates as an autonomous module performing database updates on a daily basis. It is driven by a configuration file containing the Internet addresses of database servers and paths to target files thereon, which are used to pull new material by FTP from the remote sites. If a file transfer and subsequent processing steps (optional reformatting, database insertion) are successful, the updated version of the database is added to the GeneQuiz repository.

In particular, when one of the protein or nucleotide sequence databases is updated, a non-redundant database (NRDB) of protein or nucleotide sequences is regenerated using the 'nrdb' program from National Center for Biotechnology Information (NCBI) (Gish, 1992) to filter exact replicates. The identifiers of redundant sequences from the contributing databases are retained, allowing later cross-referencing to the original databases.

Table 1 shows the databases that are currently managed by GQupdate and their size (as number of sequences) at the time of writing. Note the large reductions for non-redundant protein (down to 35%) and nucleotide (43%) database sizes, leading to economy of storage and search times.

Module: GQsearch

This module performs the basic analyses of the query protein sequence using mostly standard publicly available tools as well as some purpose-built methods (Table 2). These range over detection of motifs and biased composition regions, sequence database searching, and prediction of secondary and tertiary structural features. The system is extensible by the addition of new methods, for example, we are experimenting with new versions of sequence search methods [FASTA3 (Pearson, 1996) and PSI-BLAST (Altschul *et al.*, 1997)].

Methods may be applied directly to the query sequence or to results of previously applied methods. The collected results are then processed by the reasoning module GQreason to derive a functional characterization, and may be examined interactively using the GQbrowse module.

The set of methods is run against a query sequence in a predetermined order based on a configuration file specifying (i) dependencies between methods, (ii) command line arguments and simulated interactive input for each method, and (iii) parsers to convert each method's output to Relational Database Management System (RDB) format. The latter is a simple relational database format manipulable using the RDB tools (Hobbs, 1993). Single queries or batches of sequences (e.g. whole genome ORF sets) can be analysed, distributing runs in parallel on multiprocessor UNIX nodes or distributed over a set of UNIX workstations.

The methods can be separated into three categories according to their role in the GeneQuiz engine: **sequence filters** used to mask parts of the sequence that may adversely affect the performance of sequence database search methods; **comparison methods** that are applied to establish an automatic functional annotation; **support methods** that are run on the

Table 2. Sequence analysis tools used in GeneQuiz. These are grouped into sequence filters, used to pre-process sequences before application of comparison methods, which screen a query against sequence databases to find candidate homologues for function transfer, and support methods, which add extra sequence annotation for report generation

Description	Method	Reference
Sequence filters		
Low complexity regions	seg	Wootton and Federhen (1996)
Amino acid biased regions	biasdb	Casari & Ouzounis (unpublished)
Comparison methods		
Protein and nucleotide sequence database search	TBLASTN, BLASTP	Altschul *et al.* (1990)
	TFASTA, FASTA	Pearson and Lipman (1988)
Support methods		
Repeat prediction	repeats	M. Vingron, DKFZ, Heidelberg
Coiled-coil prediction	coils	Lupas (1997)
Blocks motif search	blimps	Henikoff *et al.* (1997)
PROSITE motif search	prosearch	Bairoch *et al.* (1997)
Multiple sequence alignment	MaxHom	Sander and Schneider (1991)
Secondary structure pred.	PredictProtein (2D)	Rost and Sander (1993)
Transmembrane helix pred.	PredictProtein (tmb)	Rost *et al.* (1995)
Residue solvent exposure pred.	PredictProtein (exposure)	Rost and Sander (1994)
3D homology modelling	WHATIF (whatif_model)	G. Vriend, EMBL, Heidelberg

sequence to provide the user with additional evidence to confirm/deny the automatic annotation.

Sequence filters. Compositionally biased (or low-complexity) regions in proteins are known to affect evaluation of the significance of database searches by identifying similar regions that are not necessarily related by divergent protein evolution. The problem has been dealt with in the past [‘seg’ (Wootton and Federhen, 1996); ‘xnu’, (Claverie and States, 1993)].

In GeneQuiz, low-complexity regions are found using seg, and amino acid-biased composition regions are detected with the program ‘biasdb’ (G.Casari and C.Ouzounis, unpublished). This performs a single-pass, ungapped comparison between a query sequence and an ideal homopolymer, identifying with a given cut-off score both the regions and the type of compositional bias with superior performance to the previous approaches (Casari *et al.*, 1996). These regions are rich in one particular amino acid that may correspond to functional or structural features of the protein, e.g. transmembrane segments or runs of charged residues.

Comparison methods. The screening of an uncharacterized query protein sequence directly against protein sequence databases or indirectly against nucleotide sequence databases by six-frame translation of the latter is common practice. Two standard search methods are used in GeneQuiz giving either ungapped local alignments [BLAST (Altschul *et al.*, 1990)] or global gapped alignments [FASTA (Pearson and Lipman, 1988)] of the query and the target sequences together with a similarity score and a significance value.

In GQsearch, a BLAST search is made, followed by FASTA if no reliable hits were found on the first pass. Bearing in mind that the primary purpose of the system is to determine function, this reduces processing time and storage requirements, important considerations when analysing a whole genome. Performance is also considerably improved by using biasdb (above) to mask amino acid-biased regions of the query, thereby reducing the incidence of false positives.

Support methods. These methods are conveniently further subdivided into pattern detection, multiple alignment and structural inference categories.

(i) *Pattern detection.* Several methods are used to scan the query sequence for repeated patterns and known motifs. Repeated sequences of amino acids are detected using the program ‘repeats’ (M.Vingron, DKFZ, Heidelberg). These often reflect large-scale features of the sequence, e.g. structural domains, and are frequent in structural proteins. Similarly, coiled-coil regions are predicted using the program ‘coils’ (Lupas, 1997).

The query is also scanned against databases of protein sequence motifs [PROSITE (Bairoch *et al.*, 1997) and ‘Blocks’ (Henikoff *et al.*, 1997)]. The presence of a motif can confirm an otherwise weak homology. Normally, these motifs correlate with functional properties, for which detailed annotations and cross-references may be available in PROSITE. The Blocks database has a more extensive collection of motifs, including those of families without known function, but is less comprehensively annotated.

(ii) *Multiple alignment.* The 'MaxHom' program (Sander and Schneider, 1991) is applied to the query and database search hits (above). MaxHom accumulates and aligns sequences to the query, most similar first, excising unaligned loops from the hits to prevent gap insertion in the query. The result is a query-centric multiple alignment for input to some of the support methods outlined below.

(iii) *Structural inference.* Several programs are able to exploit, or depend upon, the extra structural information implicit in a multiple alignment. The PHD suite of programs (Rost *et al*., 1994) makes use of the MaxHom output to produce predictions of secondary structure (Rost and Sander, 1993), transmembrane helices and connecting loop topology (internal/external) (Rost *et al*., 1995), and of residue solvent exposure (Rost and Sander, 1994).

Lastly, given a MaxHom alignment that includes a good homology to a sequence with known 3D structure, the system builds a model 3D structure of the query sequence using the WHATIF program [G.Vriend, European Molecular Biology Laboratory (EMBL), Heidelberg]. It is important to note that the model corresponds only to those parts of the query that have homology to the sequence with known 3D structure, as aligned by MaxHom.

Module: GQreason

The main purposes of the GQreason module are 2-fold: (i) to determine a broad cellular function for the query sequence family by analysing the set of homologues to the protein, i.e. to assign the family to a general functional class; (ii) to assign a specific function to the query, if possible, by transferring that function from one of the homologues. Both tasks depend on the careful choice of homologues and on the systematic analysis of sequence database annotations.

The homologue list is selected from the union of sequence hits reported by the database search programs from the preceding GQsearch stage, using only those sequences that exceed method-specific score or significance values [BLASTP, $P(N) < 1e-10$; FASTA, $score > 130$, corresponding to the 'clear' categories in Table 5].

Systematic extraction of functional information from annotations expressed in various database-specific field types (description, keyword, comment, etc.) and formats presents a harder problem. Three general criteria affecting annotation quality are:

1. *Sequence similarity.* Clearly, the higher the similarity between the query sequence and a putative homologue, the more confidence in any functional inference. This is, in turn, dependent upon the choice of database search engine or sequence comparison method chosen. GeneQuiz uses established search methods (BLAST, FASTA) with their own relative strengths, as discussed elsewhere in the literature. However, in an integrated system like GeneQuiz, there remains the issue of comparing results from disparate scoring schemes.

2. *Database quality.* The choice of databases searched is important. GeneQuiz, through the GQupdate module, ensures that a wide selection of sequence databases are accessed. However, external databases differ in the quality of curation and in the amount of annotation they offer, depending on their intended purpose. Sequence databases may contain either nucleotide (GenBank, EMBL) or protein sequences (WormPep, GenPept, TREMBL, PIR, SWISS-PROT). A quality control mechanism is required that will differentiate between annotations derived from different databases, perhaps using some explicit ranking that indicates the relative confidence of the system in each database, whether protein or nucleotide.

3. *Annotation quality.* The annotations currently found in databases are highly heterogeneous and sometimes inconsistent in the use of database fields. The provenance of a functional annotation is generally not apparent—as with database quality, a function may have been inferred by homology, or it may have been assayed experimentally (leading presumably to more reliable annotation), but this is not expressed at the annotation level in a machine-readable or even consistent manner. Annotations are generally hand crafted and inevitably reflect idiosyncrasies of the annotator despite attempts at standardization by the curators. Typical forms of description encountered include those shown in Table 3.

Automatic assignment to a functional class cannot rely on the annotations in the current generation of databases: these are inadequate as they describe protein function at a very detailed level where possible (e.g. a given sequence may be annotated as a cdc2 kinase, but not as being involved in intracellular communication). One approach is to narrow the focus to the most reliable part of any annotation, the keywords, since these may derive from or constitute a controlled vocabulary.

If a more specific functional annotation for the query than that from keyword analysis of the homologues is required, a detailed deconstruction of the text contained in the sequence annotation is required. Further, any valid system for functional transfer, whether manual or automatic, must make competent decisions on which annotation, if any, to apply to a query sequence given a list of plausible homologues and their associated annotations. Because of the complexities of, and interplay between, the three criteria of sequence similarity, database quality and annotation quality, often the best

Table 3. Typical forms of functional annotation. Most of these may be modified by descriptors (e.g. 'putative', 'by similarity'), indicating that the function was assigned on the basis of some sequence similarity, methods, parameters, or cut-offs, often unspecified. The Accept? column shows the action of the lexical analyser embodied in the functional transfer procedure in GQreason

Description	Accept?
Reference to the protein itself	
GLUTAMATE-1-SEMIALDEHYDE 2,1-AMINOMUTASE (EC 5.4.3.8) (GSA)	yes
NEGATIVE REGULATOR OF GENETIC COMPETENCE MECB	yes
Reference to other components	
penicillin-binding protein	yes
sigma-54 interacting protein	yes
Systemic information	
VIRULENCE FACTOR MVIM	yes
PROTEIN RESPONSIBLE FOR OXETANOCIN A RESISTANCE	yes
S. cerevisiae essential gene from chromosome IX, complete cds.	no
34 KD ANTIGENIC PROTEIN	no
Structural information or cellular localization	
transmembrane glycoprotein CD68, 110K - human	no
B.taurus mRNA for novel cytoplasmic protein	no
Gene name	
DnaJ, HyaA (products characterized and function known)	yes
YydK, YydF (hypothetical genes of *B. subtilis*)	no
DNA-level information	
zt19h03.r1 Soares ovary tumor NbHOT Homo sapiens cDNA clone 713621 5'	no
HYPOTHETICAL 68.5 KD PROTEIN IN SCS3-SUP44 INTERGENIC REGION	no

functional transfer may not be the one associated with the best-scoring database hit.

The solutions adopted by GeneQuiz for general functional classification and specific functional transfer are described in the next sections.

General functional class. The method used by GeneQuiz for general functional classification is based on the generation of a dictionary that associates keywords characteristic of a sequence with a set of functional classes. The keywords are as defined in the SWISS-PROT protein sequence database. GeneQuiz currently works with 14 classes of cellular function based on those of Riley (1993) and grouped into three superclasses, ENERGY, COMMUNICATION, INFORMATION (Tamames *et al.*, 1996), plus a catchall class OTHER (Table 4), although the following algorithm can be applied to any classification scheme or number of classes.

The generation of a dictionary starts with an initial comprehensive training set of example proteins classified into functional classes by a human expert. For every one of those proteins, their corresponding keywords are extracted and each is scored by the number of times that it appears in a functional class.

A filtering procedure is applied to eliminate those keywords with no functional meaning (e.g. '3D structure', 'hypothetical protein') and those that are present in just one sequence. Each one of the resulting set of keywords is assigned uniquely to a functional class if no less than 85% of its occurrences belong to that class.

Assignment of a new sequence to a class is by look-up of the keywords for that sequence in the dictionary to determine the most frequently associated class, which is then chosen. Iterative application of the assignment process to all sequences in a sequence database yields a new set of keyword/

Table 4. Functional classification used in GeneQuiz. Proteins not covered by the first three superclasses are placed in 'OTHER', which is not used in the GeneQuiz analysis

Superclass	Functional classes
ENERGY	amino acid biosynthesis biosynthesis of cofactors central intermediary metabolism energy metabolism fatty acid and phospholipid metabolism nucleotide biosynthesis transport
COMMUNICATION	cell envelope cellular processes regulatory functions
INFORMATION	replication transcription translation
OTHER	–

Table 5. Reliability values and categories. Method-specific scoring schemes are transposed onto a 0–1 scale incorporating a large bias that favours protein over nucleotide database hits. Actual reliability values used in GQreason also differentiate between protein databases by applying a further small bias to these values: SWISS-PROT, unmodified; PIR, –0.002; TREMBL, –0.004; GenPept, –0.006

Method-specific score	Reliability			
	values		categories	
	protein	nucleotide	protein	nucleotide
BLAST				
$p(N) < 1e-70$	1	0.7	clear	tentative
$p(N) < 1e-20$	0.99	0.69	clear	tentative
$p(N) < 1e-10$	0.95	0.65	clear	tentative
$p(N) < 1e-4$	0.7	0.4	tentative	marginal
$p(N) < 0.1$	0.3	0	marginal	unknown
$p(N) < 1$	0	0	unknown	unknown
FASTA				
$score > 500$	1	0.7	clear	tentative
$score > 250$	0.99	0.69	clear	tentative
$score > 145$	0.95	0.65	clear	tentative
$score > 130$	0.91	0.61	clear	tentative
$score > 90$	0.3	0	marginal	unknown
$score > 0$	0	0	unknown	unknown

class associations that can be used to generate a more extensive dictionary with an increase in classification quality (Tamames *et al.*, 1998).

In GeneQuiz, the keywords associated with the majority of all SWISS-PROT homologues of the query sequence—having suitable keyword information—are selected. Then the dictionary of keyword/class associations is used to attempt a classification of the query into one of the 14 functional classes. The collated keywords and also the species distribution of the family members are stored for later reporting in the GQbrowse module.

Transfer of specific function. GeneQuiz applies a lexical analysis procedure to the description fields of the query homologues to recognize likely functionally meaningful annotations. Currently, the system is not applied to comment fields (which are typically less structured than descriptions), nor does it take into account or try to ascertain whether or not the function has been experimentally assayed or just derived by similarity.

The GQreason module applies the following algorithmic approach to the list of homologues: (i) for each database search method, assemble a separate list of homologues, descriptions and scoring information ordered by similarity to the query; (ii) transpose method-specific scoring into a common 'reliability value' scheme which incorporates biases favouring certain databases (SWISS-PROT > PIR > TREMBL, GenPept > EMBL, GenBank) as detailed in Table 5; (iii) concatenate the lists placing favoured search methods first (BLAST > FASTA); (iv) iterate over the partially sorted list, applying a lexical analysis to each functional description, either accepting or rejecting it according to the forms shown in Table 3.

Lexical analysis consists of a series of tests for informational content (or lack of it) using first regular expressions, then known words. Referring to Figure 2, test (a) for an invalid functional description may lead to immediate rejection. Otherwise, known functionally content-free text is masked in (b), then tested for a functionally informative description in (c) and accepted accordingly. If test (c) fails, the text is further masked for non-word characters, short words under five characters, and numbers in (d) and, provided some unmasked text still remains, the description is accepted.

Of the descriptions accepted (if any) by this analysis, the one having the highest reliability value is carried over as the functional annotation of the query sequence, and the procedure terminates. The reliability value is taken as an estimate of the quality of the functional transfer, and is also transformed into a categorical scheme {'clear', 'tentative', 'marginal', 'unknown'} (see Table 5) for reporting in GQbrowse.

Module: GQbrowse

The user accesses the results of a GeneQuiz analysis via a set of HTML pages containing tabular information and graphical displays of alignments and structures, that is navigable in any table-compliant Web browser.

Pre-computed analyses of related groups of proteins, typically whole-genome ORF sets, can be examined individually by ORF identifier, or tabulated by category (e.g. functional

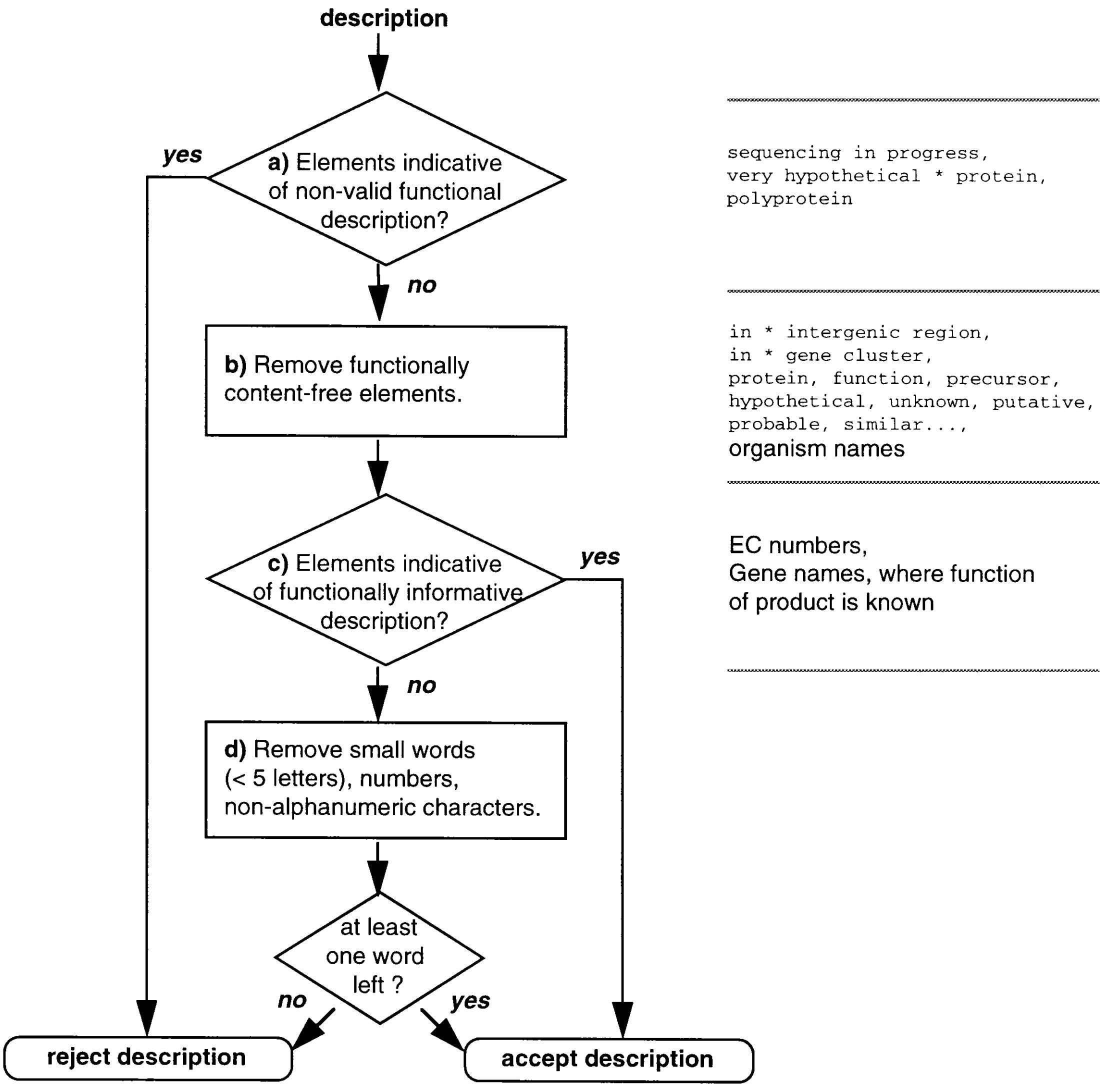

description
yes
a) Elements indicative of non-valid functional description?
no
b) Remove functionally content-free elements.
c) Elements indicative of functionally informative description?
yes
no
d) Remove small words (< 5 letters), numbers, non-alphanumeric characters.
at least one word left ?
no
yes
reject description
accept description
sequencing in progress, very hypothetical * protein, polyprotein
in * intergenic region, in * gene cluster, protein, function, precursor, hypothetical, unknown, putative, probable, similar..., organism names
EC numbers, Gene names, where function of product is known

Fig. 2. Flowchart of the lexical processing of sequence descriptions. Processing is normally by steps (a) through (d), at which point the description is accepted as functionally meaningful if, after the modifications in steps (b, d), it still contains at least one word. Shortcuts at steps (a, c) lead to immediate acceptance or rejection of the description. In practice, the flowchart may be traversed twice for each input description: an initial pass scans for known grammatical constructs using regular expression matching, and a second pass scans for special words. Examples are given alongside steps (a, b, c) and in Table 3.

class, ORF name, etc.). Summary statistics show the occupancies of such categories as well as the overall coverage of a genome in terms of reliability of functional assignment.

Findings for each ORF, as above, or for a user-supplied protein query submitted to the server, are presented in the form of a report giving structured access to and views of the collected results (see Table 2 for the list of underlying tools) from which the GeneQuiz functional assignment is inferred. The report, outlined below, is linked to a comprehensive help page detailing all section contents.

The report document comprises (i) basic information about the query sequence, such as sequence database aliases, corresponding gene names and the original functional assignment (if any) associated with the query sequence. The user can compare this with the GeneQuiz inferred functional assignment in the next section, (ii) functional information, i.e. the transferred functional assignment, the general functional class and a list of functional keywords abstracted from similar database entries determined by the searches. The reliability value of the annotation source sequence computed in GQreason is given in both numerical and categorical forms.

The functional assignment is augmented by structural information (iii) covering primary and secondary structural details: amino acid composition-biased regions (seg, biasdb), internal repeats (repeats), coiled-coil predicted regions (coils) and transmembrane helices (PHD). If a sufficiently close Protein Databank (PDB) homologue was found, a homology-built tertiary structure (WHATIF) can also be viewed in an external viewer, e.g. RASMOL (Sayle and Milner-White, 1995). The phylogenetic range of the sequence family (iv) is indicated by species and taxa membership lists extracted from the underlying database searches. For this purpose, a dictionary of species and taxa for all sequences in the NRDB is generated with each update, using only those species and taxa names found in SWISS-PROT, and excluding artificial sequences. Finally, a section (v) detailing the search results and statistics allows the user to examine the raw sequence search listings (BLAST, FASTA, MaxHom), motif search results (PROSITE, Blocks) and a merged table of all reliable homologues by the different search methods.

Throughout the report, database sequence identifiers and PROSITE and Blocks entries are hyperlinked through SRS to the original database entry. As well as the tabular data, many of the sequence annotations and the sequence alignments implied by the search methods may be viewed graphically. At several points in the report, links may be followed to display the linear features (composition bias, coiled-coil, secondary structure and transmembrane predictions, motif positions) aligned against the query, with embedded links in these features allowing interrogation of the specific feature data for that region (where appropriate). Similarly, database search results (BLAST, FASTA) or multiple alignments (MaxHom) can be viewed graphically as coloured alignments with links via SRS to the original sequences.

These graphical presentations depend on the MView software and libraries (Brown *et al.*, 1998). MView is a tool for converting the results of a sequence database search or multiple alignment into an HTML page showing a coloured alignment. The example shown in Figure 3 illustrates the colouring scheme applied by MView which is based on identity with the query sequence and amino acid properties. Use of these displays greatly facilitates visual interpretation of search results by highlighting conserved regions, even when score or significance values are extremely weak, and by showing their correspondence (or otherwise) with known motifs and predicted structural features.

Results and discussion

The GeneQuiz system is designed to analyse a single query protein sequence, or a batch of sequences such as a set of translated ORFs from a sequencing project, to (i) assign where possible a function and (ii) collate various information derived using different analytical and predictive tools.

It should be stated clearly that this kind of system is not a substitute for expert analysis. Rather, it frees the user from the tedious and repetitive tasks of searching and collating, allowing more time to be devoted to the important stage of expert manual verification of the results.

There are a number of features of the system that make it practicable and useful, offering advantages over unassisted manual functional annotation. These are: (i) automation, with the corollaries that analysis methods are applied in an objective and repeatable fashion; (ii) use of multiple, up-to-date sequence databases; (iii) use of multiple homology search methods; (iv) provision of on-line browsing tools for examining results.

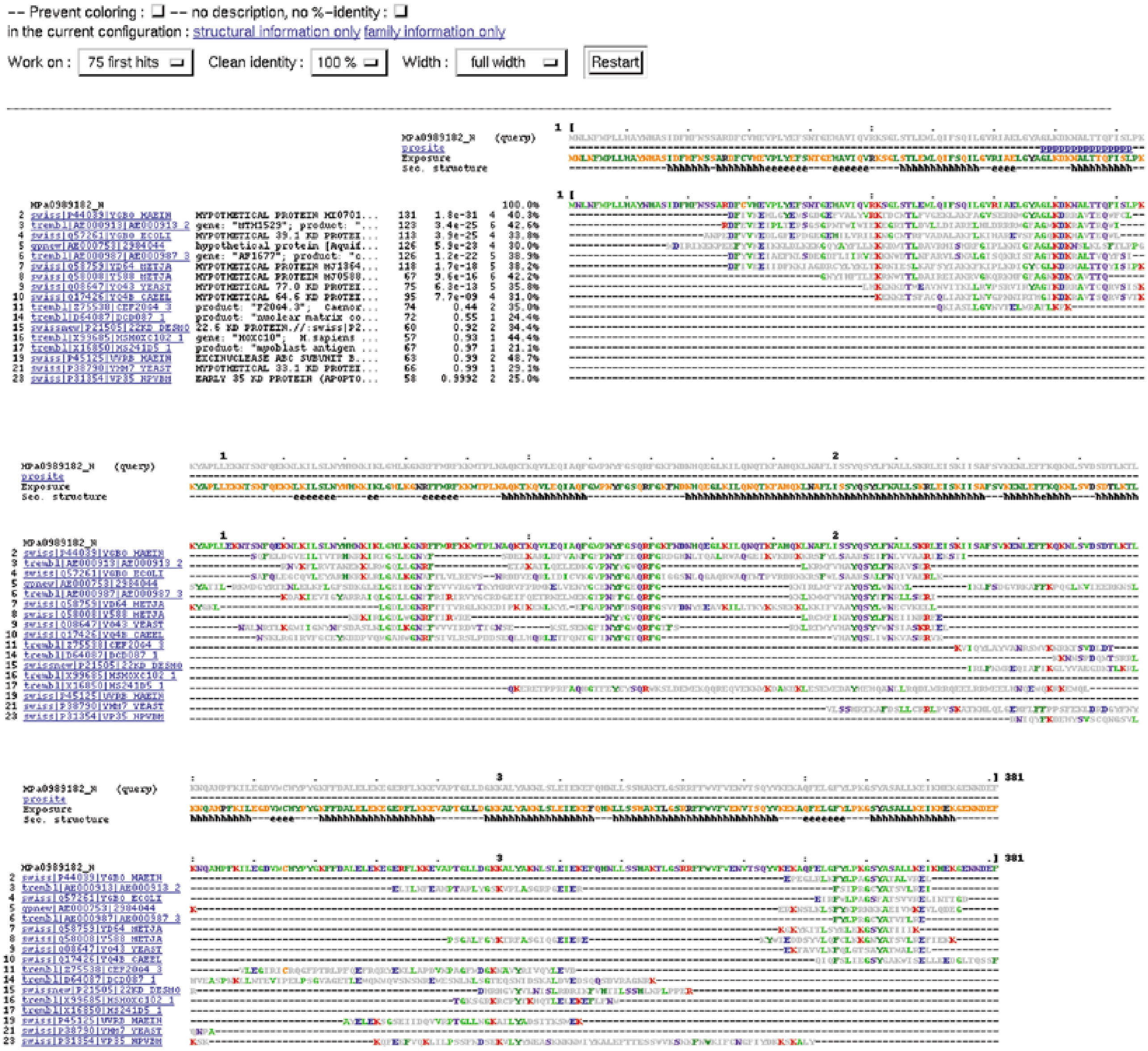
Perhaps the most obvious feature of GeneQuiz is that it is automatic. This is a prerequisite for the consistent functional characterization of the very large numbers of protein sequences deriving from genome sequencing projects. Table 6 lists the genomes that have been analysed to date using GeneQuiz. Automation also permits the same system to be re-applied to a set of ORFs as the sequence databases mature. The objectivity and consistency of operation of the underlying analyses and report synthesis facilitate the analysis of large groups of proteins [gathered by family (García-Ranea and Valencia, 1998), or relation to diseases (Andrade *et al.* 1998)], fragments of genomes (Voss *et al.*, 1995, 1997) and complete genomes (Casari *et al.*, 1995b; Ouzounis *et al.*, 1996b; Andrade *et al.*, 1997), and the comparison of results between different genomes (Ouzounis *et al.*, 1996a; Tamames *et al.*, 1996; Andrade *et al.*, 1999).

The major use of GeneQuiz has hitherto concentrated on whole-genome analysis, since, apart from academic interest, these datasets are the largest source of uncharacterized protein sequences, and therefore the best training set for a system for automatic functional annotation of proteins.

Query sequences supplied as ORFs from genome projects are usually already annotated by the sequencing group. As with database functional annotations, the quality of this information varies from very specific definitions to functionally weak, or content-free, descriptions, such as ‘unknown’, ‘hypothetical protein’, ‘conserved protein similar to …’. Regardless of quality, the sequencing group raw annotations are ignored by GeneQuiz for the purpose of assigning function.

Eventually, the ORFs are deposited in a public database by the sequencing group and then percolate through various database membership and curation stages. The surviving curated annotations now become available for GeneQuiz to assign, so that all functional annotations transferred by GeneQuiz are either taken directly from a database sequence identical to the query (perhaps already deposited), or else transferred from a similar database sequence having a valid annotation (as defined in GQreason).

The sequencing group raw ORF annotations are considered, however, when determining new findings in GeneQuiz. There are three criteria that must be satisfied for a GeneQuiz automatic functional transfer to qualify as a new finding: (i) the transfer must have been made with reliability ‘clear’ (see Table 5); (ii) the sequence must not be identical to the database sequence



-- Prevent coloring : □ -- no description, no %-identity : □
in the current configuration : structural information only family information only

Work on : 75 first hits   Clean identity : 100 %   Width : full width   Restart

```
MPa0989182_N
 2 swiss|P44039|YGBO_HAEIN        HYPOTHETICAL PROTEIN HI0701...   131   1.8e-31  4  40.3%
 3 trembl|AE000913|AE000913_2     gene: "MTH1529"; product: "...   123   3.4e-25  6  42.6%
 4 swiss|Q57261|YGBO_ECOLI        HYPOTHETICAL 39.1 KD PROTEI...   113   3.9e-25  4  33.8%
 5 gpnew|AE000753|2984044         hypothetical protein [Aquif...   126   5.9e-23  4  30.0%
 6 trembl|AE000987|AE000987_3     gene: "AF1677"; product: "c...   126   1.2e-22  5  38.9%
 7 swiss|Q58759|YD64_METJA        HYPOTHETICAL PROTEIN MJ1364...   118   1.7e-18  5  38.2%
 8 swiss|Q58008|Y588_METJA        HYPOTHETICAL PROTEIN MJ0588...    67   9.6e-16  6  42.2%
 9 swiss|Q08647|Y043_YEAST        HYPOTHETICAL 77.0 KD PROTEI...    75   6.3e-13  5  35.8%
10 swiss|Q17426|Y04B_CAEEL        HYPOTHETICAL 64.6 KD PROTEI...    95   7.7e-09  4  31.0%
11 trembl|Z75538|CEF2064_3        product: "F2064.3";  Caenor...    74      0.44  2  35.0%
14 trembl|D64087|DCD087_1         product: "nuclear matrix co...    72      0.55  1  24.4%
15 swissnew|P21505|22KD_DESMO     22.6 KD PROTEIN.//:swiss|P2...    60      0.92  2  34.4%
16 trembl|X99685|HSMOXC102_1      gene: "MOXC10";  H.sapiens ...    57      0.93  1  44.4%
17 trembl|X16850|HS241D5_1        product: "myoblast antigen ...    67      0.97  1  21.1%
19 swiss|P45125|UVRB_HAEIN        EXCINUCLEASE ABC SUBUNIT B....    63      0.99  2  48.7%
21 swiss|P38790|YHM7_YEAST        HYPOTHETICAL 33.1 KD PROTEI...    66      0.99  1  29.1%
23 swiss|P31354|VP35_NPVBM        EARLY 35 KD PROTEIN (APOPTO...    58    0.9992  2  25.0%
```

Fig. 3. Alignment of a new family found using GeneQuiz. The upper section of each pane shows sequence and structure annotations: in th is case PROSITE patterns (‘prosite’), predicted solvent exposure (‘Exposure’) and predicted secondary structure (‘Sec. structure’) aligned against the query, while the lower section, which is in register, shows a composite alignment of similar fragments to the query sequence found by a BLASTP database search and generated using MView. The colour scheme for the ‘Exposure’ row is: orange, exposed; green, buried. Symbols for the ‘Sec. structure’ row are: ‘e’, β-strand; ‘h’, α-helix. Residues in the sequence alignment are coloured by identity to the query and colour coded by physicochemical property: greens, hydrophobic; blue, negative charge; red, positive charge; purple, polar; orange, cysteine. In the BLASTP-derived alignment, labels for each sequence at left give the search rank, database identifier (linked to the database entry via SRS), and BLASTP *P*-value. Columns or blocks of coloured residues and corresponding annotations highlight conserved patterns indicative of family membership. In this example, the *H.pylori* query sequence (predicted ORF 989182 to 990324 on the default strand) has five regions with patterns that appear to be conserved in the first 10 sequences (thus likely to form a family), but not in the remaining sequences, which are shown for illustration. Note that the last obvious family member (rank 10) has a BLASTP value of 7.7e – 9, below the automatic threshold for clear BLASTP hits used by GQreason (see System section), showing the importance of visual examination of results. This family is predicted by PHD to be mainly α-helical (on the basis of the underlying MaxHom alignment), and seems not to contain any coiled-coil, transmembrane or low-complexity regions, which would otherwise have been reported as extra rows in the annotations. The N-terminal region contains a match to PROSITE pattern PS01268 (uncharacterized protein family UPF0024 signature; unpublished observations, A.Bairoch, 1997), which is shown in the ‘prosite’ row as a live link (blue ‘p’s) to that PROSITE entry via SRS. Although all members are as yet hypothetical proteins, the family is clearly real and, spanning members of archeal, bacterial and eukaryotic kingdoms, is presumably both ancient and fundamentally important.

Table 6. Complete genomes and GeneQuiz automatic functional assignment statistics. 3D, sequences for which a model could be built by similarity to a sequence of known 3D structure; F, sequences for which function is known or can be inferred from similar sequences; S, sequences that have similar sequences in the databases. Total throughput for the 12 genomes: 37 Mb or 29 145 ORFs, with average levels of functional assignment: 13% 3D, 59% F, 84% S

Organism	Sequencing reference	Size		3D	F	S	Run
		Mb	ORFs	%	%	%	date
Bacteria							
Haemophilus influenzae Rd	Fleischman *et al.* (1995) Science	1.8	1680	17	71	99	2/98
Mycoplasma genitalium	Fraser *et al.* (1995) Science	0.6	468	15	70	87	6/97
Mycoplasma pneumoniae	Himmelreich *et al.* (1996) NAR	0.8	677	10	59	93	12/96
Synechocystis sp.	Kaneko *et al.* (1996) DNA Res.	3.6	3168	12	56	74	11/97
Escherichia coli	Blattner *et al.* (1997) Science	4.7	4285	14	72	87	4/97
Helicobacter pylori	Tomb *et al.* (1997) Nature	1.7	1590	11	54	84	8/97
Bacillus subtilis	Kunst *et al.* (1997) Nature	4.2	4100	16	62	83	1/98
Borrelia burgdorferi	Fraser *et al.* (1997) Nature	1.4	850	15	64	82	1/98
Archaea							
Methanococcus jannaschii	Bult *et al.* (1996) Science	1.7	1735	8	45	74	10/96
Archeoglobus fulgidus	Klenk *et al.* (1997) Nature	2.2	2437	11	47	75	12/97
Methanobacterium thermoautotrophicum	Smith *et al.* (1997) J. Bacteriol.	1.8	1871	11	53	89	11/97
Eukarya							
Saccharomyces cerevisiae	Goffeau *et al.* (1996) Science	12.5	6284	11	60	77	10/97

from which the annotation was taken; (iii) the original raw annotation supplied by the sequencer must be known to lack functional information.

The first criterion is an obvious quality control. The second arises because, as noted above, a raw sequence eventually finds its way into one of the databases mined by GeneQuiz to determine function; findings based on self are discounted. Of course, the function is still transferred by GeneQuiz to the query, but that transfer cannot be considered as a new finding as it is simply a copy of a previously assigned and curated function. The third criterion is based on a manual assessment of the raw ORF annotations, rejecting such items if they are functionally content free: only annotations rejected in this way can contribute to the new finding list.

Each criterion errs on the side of caution. Only the most conservative of candidate new findings are noted thereby and highlighted for immediate attention or reported in performance statistics. Other annotations derived by GeneQuiz, even if not considered new findings, may complement the analysis of the sequencing group by (i) supporting the original annotation or (ii) suggesting a better alternative. Note that automatic detection of the latter situation is not attempted, because deeper expert analysis may be needed to discern which of the conflicting annotations is more accurate.

In any case, the report always displays the original raw annotation for comparison with that derived by GeneQuiz, and the user can use the browsing facilities to inspect the quality of the automatic annotation or to explore the collected results of the different analyses, which may lead to other discoveries (e.g. families of hypothetical proteins, remote homologies). New annotations transferred by GeneQuiz may ultimately be incorporated into the databases.

Again, it is important to emphasize that the transfer of annotations performed by GeneQuiz is completely independent of the original raw annotations supplied with any ORFs. By contrast, the permitted scope for possible new findings is specifically limited to include only those ORFs that lack functionally meaningful raw annotation.

Some examples of new findings inferred using GeneQuiz are presented in the rest of this section. In general, the system is able to infer a remarkable number of new functional annotations reliably. The reasons for this are discussed further in relation to the choice of databases and search methods. Some general problems or caveats that affect this kind of procedure (whether embodied in an automatic system or manually applied) are presented with suggestions for future work.

New findings

Full sets of new findings for particular genomes have been released and published elsewhere (Casari *et al*., 1995b; Ouzounis *et al*., 1996b; Andrade *et al*., 1997). Here, we illustrate the application of GeneQuiz using three selected examples from analyses of several whole genomes (*Synechocystis* sp., *Helicobacter pylori*, *Saccharomyces cerevisiae*), which identify (i) a new function, (ii) a new family and (iii) a new superfamily, respectively.

New function. In the analysis of the *Synechocystis* sp. genome, GeneQuiz identified a putative *cis*-aconitase gene (ORF 'slr0665'), which was not reported as such by the research group that sequenced the genome (Kaneko *et al*., 1996). Since *cis*-aconitase is a key enzyme of the citrate cycle, this is an important finding. GeneQuiz found clear evidence for this identity: strong sequence similarity to other bacterial *cis*-aconitases and two typical motifs of the family (as defined by BLOCKS).

GeneQuiz transferred the annotation from the *Escherichia coli* biochemically characterized *cis*-aconitase (Fujita *et al*., 1994). The database entry SWISS-PROT:ACO2_ECOLI (the notation used for database entries is database:identifier) was generated in June 1994 and updated in November 1997 to account for the correction of a frameshift. It is likely that the analysis by the original sequencers missed the similarity because of this. New similar sequences from *M.jannaschii* and *H.pylori* confirm the homology. Because GeneQuiz uses an up-to-date compound sequence database, all recent database changes are taken into account, such as the correction of the *E.coli* sequence and the inclusion of new homologues.

Other non-automatic systems for protein comparison [COG (Tatusov *et al*., 1997); WIT (Overbeek *et al*., 1999)] arrived at the same conclusions. Finally, in July 1998, the sequence was included in the SWISS-PROT database (SWISS-PROT:ACO2_SYNY3) annotated accordingly as *cis*-aconitase.

New family. The GQreason module selects a set of sequences similar to the query sequence as reliable homologues. If none has a characterized function, the module cannot assign a function to the query. However, even in this situation, GeneQuiz can provide other information to characterize the family, such as the presence of sequence motifs and a tentative indication of the taxonomic span.

Manual examination of the alignment of the query to all similar sequences (as displayed by MView) can help the user in the task of family characterization by highlighting common sequence features, even beyond the safe thresholds applied by the automatic reasoning module. Other more elaborated alignment programs can then be used to construct more rigorous sequence alignments [e.g. using CLUSTAL (Higgins *et al*., 1996) or SAGA (Notredame and Higgins, 1996)].

The discovery of new families of hypothetical proteins becomes more likely with the increasing number of hypothetical ORFs from large-scale sequencing projects. Conservation within the family validates the inference of the hypothetical ORFs, since the conservation of the translation product across species is normally an indication of their expression as functional proteins. Moreover, the alignment pinpoints conserved patterns and may indicate functionally important sites. Eventually, remote homologues with known functionality may be detected by searching the databases with the profile of the family [e.g. MoST (Tatusov *et al*., 1994), HMMer (Eddy *et al*., 1995), WiseTools (Birney *et al*., 1996)], a procedure usually more sensitive than the single sequence-to-sequence comparison methods employed by GeneQuiz in the database searches. The taxonomic span of the family may be related to the evolutionary origin of the function associated with it. A broad distribution with conservation among very divergent taxonomic branches may indicate basic functions important to the survival of the organism. Alternatively, unique occurrence in a single genus may indicate a function particular to that genus, conferring, for example, pathogenicity.

An example of the discovery of a new family of hypothetical proteins is shown in Figure 3. The closest homologues to a query hypothetical protein sequence from *H.pylori* (sequences 1–10) have a highly significant *P*-value reported by BLASTP, and can be considered to be members of the family. Conserved patterns accumulate in five blocks. Interestingly, the family spans eukaryotes (*S.cerevisiae*, *Caenorhabditis elegans*), archea (*M.jannaschii*) and bacteria (*H.influenzae*, *E.coli*). Examination of the original database entries revealed that all homologues were annotated as hypothetical proteins, and that their functions could not be predicted by similarity to other proteins of known function. Other less related sequences (sequences 11–25) have a *P*-value > 0.4, indicating very low sequence similarity. Examination of the alignment of these sequences to the query sequence shows that the similarity reported by BLASTP is not indicative of functional similarity, but reflects more general structural similarity through patterns of conserved negatively charged amino acids.

This information has also been generated by posterior non-automatic analyses (Tatusov *et al*., 1997; Doerks *et al*., 1998). The functionality of the family remains unknown.

New superfamily. The functional assignments made by GeneQuiz have an associated reliability (Table 5). In the preceding paragraphs, we have only considered 'clear' functional assignments selected by the GeneQuiz system using very conservative thresholds. Nevertheless, other less reliable sequence relationships, although they may be classified by the system as 'tentative' or even 'marginal', may have a biological basis. The system cannot resolve such cases automatically because of the risk of introducing many false positives into the putative family. However, the user, if particularly interested in a specific protein, has available all the information derived by GeneQuiz and can manually validate candidate homologies to remotely related genes and superfamilies.

One example is the hypothetical yeast ‘YCL008C’ gene (TREMBLnew:SCCHRIII_60) for which GeneQuiz transfers function from the mouse and human ‘TSG101’ genes associated with tumour susceptibility (TREMBL: MM52945_1 and TREMBL:HSU82130_1). Since the similarity levels to both proteins are well under the restrictive thresholds applied in the reasoning module (BLASTP *P*-value of 0.044, and FASTA score of 95, for the closer mouse sequence), the annotation was reported as ‘marginal’. Nevertheless, examination of the alignment indicates that the similarity is meaningful and suggests an even more distant relationship to members of the ubiquitin-conjugating protein family (SWISS-PROT:UBC4_CANAL, SWISS-PROT: UBC4_YEAST). A model of the query sequence can be built using the information from the 3D structure of one of the ubiquitins (PDB:2UCE, corresponding to UBC4_YEAST). Recently, several authors (Koonin and Abagyan, 1997; Ponting *et al*., 1997; Sancho *et al*., 1998) have independently described this similarity after extensive sequence analysis.

Database coverage

As the knowledge stored in biological databases constantly increases and changes, any analysis of ORF sets may have to be repeated. New information for a matching ORF may be gleaned under any of the following conditions: (i) a new structure is deposited in PDB permitting modelling by homology; (ii) an extant database sequence initially lacking suitable annotation is updated with functional information (either by experiment or by homology); (iii) a completely new functionally characterized sequence enters the databases; or (iv) a completely new, but uncharacterized, sequence enters the databases, which nevertheless validates by similarity the existence of a predicted ORF. The last case is often to be expected with the sequencing of phylogenetically close organisms having many common loci, e.g. *Mycoplasma genitalium* and *Mycoplasma pneumoniae* (Himmelreich *et al*., 1997).

The multiplicity of databases searched is also important. The inclusion of PDB in the GeneQuiz database has just been cited above in the context of structural inference. More generally, GeneQuiz uses a combination of protein sequence and translated nucleotide databases with which to achieve maximal coverage of the available potential functional annotations.

Thus, the regular update of the analysis of ‘old’ genomes with comprehensive and up-to-date databases continually generates new results, and, again, this can be easily done with the help of an automatic system. The effect is clearly shown by a time series of GeneQuiz-derived annotation categories for the *H.influenzae* genome (Figure 4), in which there is steady growth of the different categories of homologue, most importantly, for inferred 3D homology and for functional assignments by clear homology.

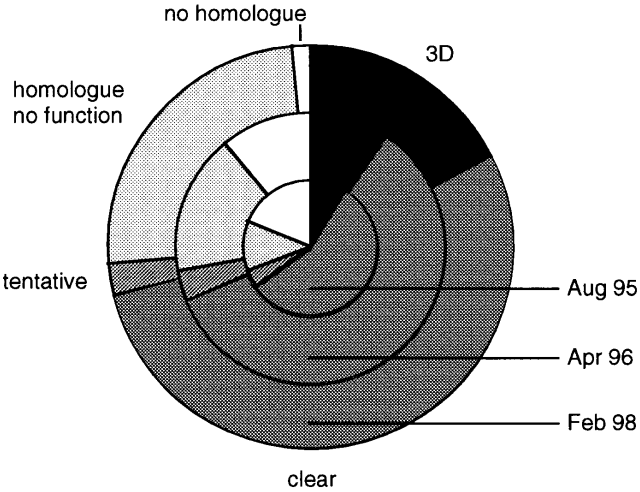


Fig. 4. Time series of GeneQuiz analyses for the *Haemophilus influenzae* genome (Fleischmann *et al*., 1995). The nested piecharts show categories of reliability of functional annotation changing with time, from inside out: August 1995 (some days after the public release of the genome), April 1996 and February 1998. Reliability levels (clockwise) are (1) with clear annotation and 3D structure predicted by homology modelling; (2) other sequences with clear annotation; (3) sequences with tentative annotation; (4) sequences with a homologue but no function; (5) sequences with no homologue. After the last analysis, only 22 sequences remained in the ‘no homologue’ class, assignment to which is very lax, since the automatic levels of similarity accepted by GeneQuiz as indicative of homology and membership of the other categories are very conservative.

For example, of the 150 annotations corresponding to the new findings reported by GeneQuiz for *H.influenzae* in August 1995 (Casari *et al*., 1995b), examination of the corresponding database entries in February 1998 showed that a total of 42% were incorporated in the database protein description [either as a sure function (27%), or as homologue (10%) or probable (5%)]. Another 56% were still defined as ‘hypothetical proteins’ (with a remarkable 44% for which the similarity to the protein selected by GeneQuiz for function transfer was indicated). A remaining 2% of annotations corresponded to ORFs that no longer exist in the databases (e.g. due to frameshift correction). One example is ORF HI0169 that turns out to correspond to the terminal part of a single ORF, now called ORF168/69. Interestingly, for some 6% of the August 1995 new findings, GeneQuiz now gives a different annotation, in most cases from a closer homologue that was not present in the databases at the time of the original run.

The origin of new findings in relation to database dynamics was examined using four recent GeneQuiz runs ex-

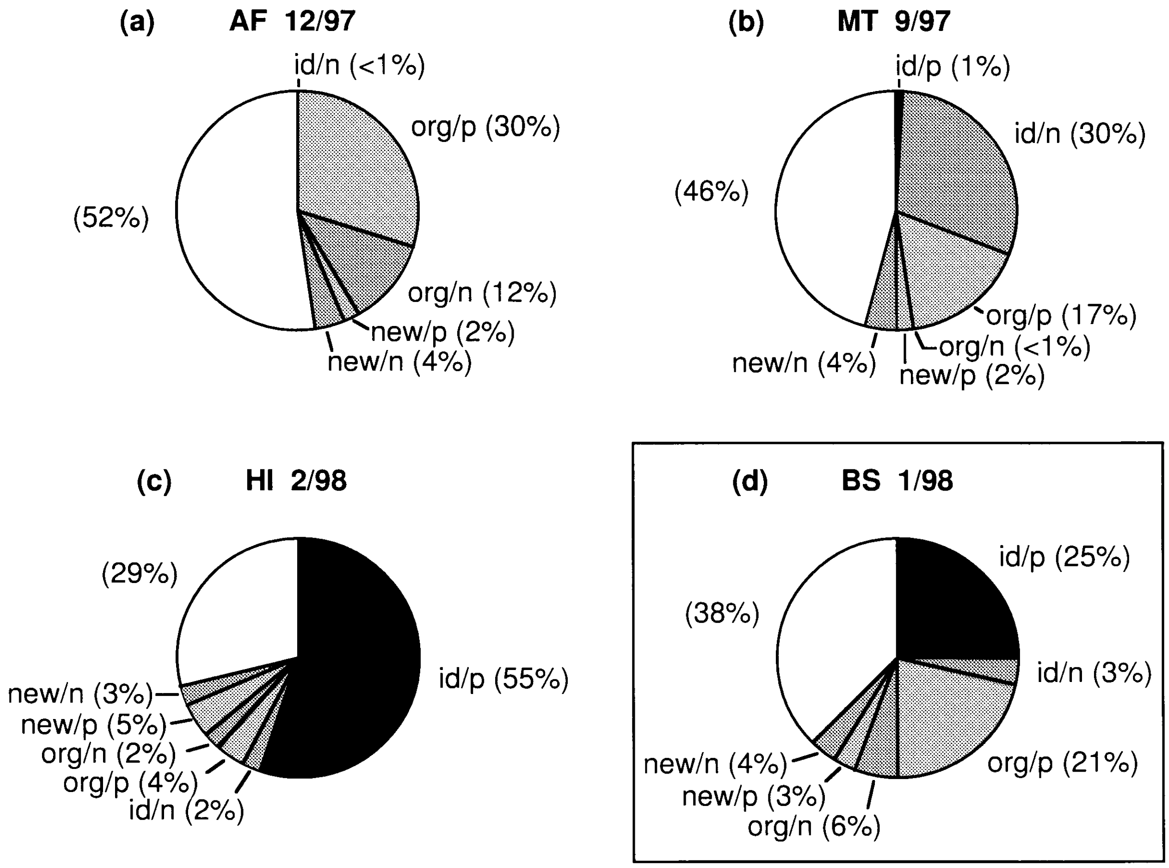


Fig. 5. Sources of functional annotation for GeneQuiz analyses of four complete bacterial and archaeal genomes. Piecharts depict the p ercentage of 'clear' ORF functional assignments deriving from the component databases of the GeneQuiz system. Each is labelled at the top with an abbreviated organism name and the date of that GeneQuiz run. (**a**) *Archeoglobus fulgidus* (AF) a few days after publication of the genome (Klenk *et al.*, 1997) (2383 ORFs); (**b**) *Methanobacterium thermoautotrophicum* (MT) some weeks after publication (Smith *et al.*, 1997) (1871 ORFs); (**c**) *Haemophilus influenzae* (HI) ~3 years after publication (Fleischmann *et al.*, 1995) (1717 ORFs); (**d**) *Bacillus subtilis* (BS) some days after publication of the full genome (Kunst *et al.*, 1997) (4099 ORFs). The clear assignments are classified: id, identical to a database sequence; org, similar to a database sequence, accepting sequencing group's original annotation; new, similar to a database sequence, rejectin g sequencing group's and assigning new annotation. These classes are subdivided by type of originating database: p, for protein databases (SWISS-PROT, PIR), n, for translated nucleotide (TREMBL, GenPept, WormPep).

hibiting various periods of delay after genome publication and widely ranging levels of experimental knowledge on the organism in question (see Figure 5).

The first three GeneQuiz analyses (Figure 5a–c) are of genomes for which no (or very few) sequences were present in any database before their publication. At the time of the runs, these three genomes ranged over (a) essentially no sequence in any database (minimal identical hit sectors; id), (b) sequences deposited only in translated nucleotide databases (large nucleotide identical hit sector; id/n) and (c) sequences deposited in protein databases (very large protein database hit sector; id/p). The idealized progression (a–c) reflects increasing database penetration.

The fourth analysis (Figure 5d, boxed) is of a long-studied bacterium for which many sequences were already annotated in protein databases years before the completion of sequencing. Yet, many new sequences have now been published that have not had time to percolate through the databases; it does not fit neatly into the progression because it is a hybrid case. The ideal application of a GeneQuiz-like system would be to genomes like (a) in the figure, for which little is known, and for which maximal new functional annotation could be inferred *de novo*.

Each piechart shows that new findings comprise 6–8% of the total number of ORFs for that genome, even the *H.influenzae* genome (c), for which ~3 years have elapsed between the publication of the genome and the depicted GeneQuiz run. To see which databases are providing the new findings, consider runs (a) and (d), which were performed shortly after publication of the whole genomes of *Archaeoglobus fulgidus* and *Bacillus subtilis*, respectively. Comparing the GeneQuiz new findings sectors (new) with the non-identical homologue sectors of original sequencer annotations (org) by the ratio of annotations deriving from translated nucleotide databases over protein databases (n/p), the ratio is much higher for the GeneQuiz new findings. This indicates that GeneQuiz more often makes discoveries based on similarities to sequences in translated nucleotide databases than do the original annotators.

This is consistent with the observation that, depending on database submission procedures, curation standards and update frequency, new data appear first in the nucleotide databases (EMBL, GenBank), then in the machine translation databases (TREMBL, GenPept) and finally in the protein databases (SWISS-PROT, PIR). There is a clear benefit in supplementing the protein databases with translated nucleotide databases when determining function by homology. Likewise, the ratio of findings based on the 'new' databases containing incremental updates (SWISSnew, TREMBLnew, GenPeptnew) versus the released databases (SWISS-PROT, TREMBL, GenPept) shows that these are the source of a large proportion of functional transfers (data not shown).

Database search methods

As described earlier, GeneQuiz uses two standard methods, BLAST and FASTA, applying the latter only when the former does not find a reliable hit. In all four genomes of Figure 5, the number of 'clear' annotations deriving from BLAST hits always exceeds the number of FASTA hits (4518/1398 or 3.2), as expected since it is run first. However, the number of FASTA-derived annotations is a large fraction of the overall count (about a quarter) showing the complementary nature of the methods, given the GeneQuiz reliability thresholds (Table 5).

Consider again the two genomes that were analysed shortly after publication (Figure 5a and d), this time examining the ratio of BLAST-initiated versus FASTA findings in the original annotation (org) and new findings (new) sectors. For *A.fulgidus*, the ratios are 6.6 and 1.3, respectively, while for *B.subtilis* they are 6.7 and 2.0. Compared to the overall ratio, they show that the original annotations derive mainly from database homologues readily identified using BLAST, while for the GeneQuiz new findings, a larger than average fraction are associated with the use of FASTA.

This is not to say that either method is superior. The complementary performances probably arise from twilight cases of homology where BLAST and FASTA scores lie either side of the GeneQuiz reliability through thresholds (Table 5) for those methods. As any similar database search scheme must rely on the use of such thresholds, this is a general problem; the use of more than one method is indicated.

It is interesting to note that there are cases in which a new finding is obvious by either method and it is unclear as to why it was overlooked by the sequencing group performing the original annotation. GeneQuiz, as an automatic system, is exhaustive and applies the same battery of methods and logic to all the proteins under analysis; there is no possible error by omission.

System limitations and further work

The many advantages of an automated system for the transfer of functional annotation from database sequences to a query sequence, including avoidance of errors such as omission, have been discussed above. However, any manual or automatic procedure is prone to several types of error (Bork and Bairoch, 1996; Bork and Koonin, 1998; Galperin and Koonin, 1998). Some of them are not currently addressed in GeneQuiz. The most important of these are: (i) false positives, where a transfer is made on the basis of a wrongly inferred homology; (ii) innacurate transfer, where the wrong information is transferred although the homology is correct; (iii) transfer of inaccurate information, where the database source is itself misleading.

We estimated an erroneous 5% of 'clear' GeneQuiz predictions in the analysis of *M.genitalium* (Ouzounis *et al*., 1996b) and *M.jannaschii* (Andrade *et al*., 1997), or even less in more recent versions of the system (2.5% estimated in the analysis of *H.pylori*; Reich *et al*., in preparation). Independent assessment by others either gives similar estimates (Kyrpides *et al*., 1996) or suggests higher error rates (Koonin *et al*., 1997; Galperin and Koonin, 1998). This is still a controversial issue since the error assessment depends on the information present in the database at the time of the estimation, the degree of belief of the experts in twilight zone findings and the care taken during the tedious manual check.

At present, we again stress that only the experienced user can resolve these problems, hence the necessity of manual scrutiny of GeneQuiz output via the browsing facilities to arrive at an expert adjudication. Even then, resolution of difficult cases may only be possible pendant upon the arrival of further correct or more detailed information in the sequence databases. The three sources of error just outlined are discussed below with examples and suggestions for future work including ongoing developments.

False positives. There are cases in which the sequence similarity detected by the database search methods correlates not with functional similarity (e.g. an active site) with implied homology, but with structural similarity which may not indicate homology. This is the case for regions of amino acid bias, transmembrane stretches rich with hydrophobic residues, and coiled-coil regions, for which, in the absence of other similarities, the function of the similar sequence should not be transferred to the query.

For example, the *Borrelia burgdorferi* gene 'BB0071' (TREMBLnew:AE001120_4), annotated by the original sequencers as 'hypothetical protein' (Fraser *et al*., 1997), had function assigned by GeneQuiz with 'clear' similarity to the

'NADH-ubiquinone oxidoreductase chain 2 (EC 1.6.5.3)' from *Paramecium tetraurelia* (SWISS-PROT:NU2M_ PARTE), a transmembrane protein located in the inner mitochondrial membrane. The combined view of the PHD output with the BLASTP traces using MView showed that the alignment (query sequence positions 141–282; FASTA score just above the 'clear' cut-off) corresponded to transmembrane stretches, and was not consistent with the conservation patterns for the query with other similar sequences.

GeneQuiz already filters sequences by masking regions of amino acid bias so that they are ignored by the database search methods and non-homologous matches of this type are excluded from the database searches. However, as the above example shows, more such filtering is required. A more general strategy would combine various sources of information (secondary structure, transmembrane, coiled-coil predictions, etc.) to deduce a consistent map of the query sequence (and derived alignments) in terms of regions of low or high information content, using only the latter for homology and thence functional inferences. Of course, any such system would also be liable to over-masking of good regions, necessitating a careful balance.

Inaccurate transfer. Given a valid homology, the GeneQuiz system assigns function by the transfer of the description of one, and only one, database sequence to the query. This simplistic approach is effective (as shown above) and is easy to follow and control. However, there are cases for which this procedure is unable to discriminate correctly between possible functions for proteins comprising multiple functional domains, or for proteins that are members of closely related subfamilies. Two examples illustrate the domain problem and the functional hierarchy problem, respectively. In both cases, solutions may involve the analysis of not one, but of multiple homologues.

(i) *Domain problem.* The function of a protein is usually associated with, indeed is defined by, certain structural regions, such as surfaces of interaction and active sites, and their physicochemical properties. Many proteins comprise multiple functional units, each associated with perhaps a different function, although the assemblage may have some higher level compound role as a consequence of the interaction of the domains through their structural adjacency.

GeneQuiz transfers function from homologue to query, considering each as indivisible units with a singular function. Sources of error from such a naive treatment relate to the potential disparity between the regions conferring function and those matched by the database search method. Possible errors are (a) transfer of function from homologue to query via an unassociated region of the homologue sequence and, conversely, (b) failure to identify the function of a region of the query that remains unmatched by the selected homologue.

For example, in the GeneQuiz analysis of the *M.genitalium* genome, the hypothetical protein MG449 (SWISS-PROT:Y449_MYCGE) was found to be similar to the β-chain of the *E.coli* phenylalanyl-tRNA synthetase (SWISS-PROT:SYFB_ECOLI). On examination of the alignment, the similarity was found to be constrained to a region of ~160 amino acids at the N-terminal of the homologue, leaving >600 C-terminal amino acids of the homologue unmatched [problem (a)]. Although the region of similarity is found in bacterial β-chain phenylalanyl-tRNA synthetases, it has been demonstrated that it forms a domain common to other proteins of various function (see the discussion by Koonin *et al*., 1997). Also, the query is not entirely covered by this domain, with >100 N-terminal amino acids remaining unmatched [problem (b)], so the automatic annotation, although based on true similarity, is incorrect or insufficiently specific.

The approach to the first problem requires that more information be gathered on the correlation of function with sequence location for a given homologue. This information could be determined experimentally and annotated in the corresponding database entry, or it may be derived automatically from a multiple sequence alignment of the family: a well-conserved region among a statistically significant number of proteins could be considered a valid domain, with function inferred by consensus from the containing homologues. A solution to the second problem requires the analysis of other homologues to find a possible match for the unassigned region of the query.

(ii) *Functional hierarchy.* Within a protein family, there may be specializations of function within an overall general class. For example, the Enzyme Commission (EC) system attempts to organize enzymes into a hierarchical scheme based on types of reaction catalysed. A given evolutionarily related family of enzymes may contain proteins catalysing a similar reaction, but with different substrate specificities. This situation can lead to overprediction (too specific) or underprediction (too general) of function due to inappropriate choice of the family member from which to transfer function.

For example, the *B.subtilis* gene 'yesQ' belongs to a family of proteins that translocate a substrate across the membrane, forming part of a bacterial transport system. More specifically, it belongs to a subfamily that imports sugars. This gene was cautiously annotated by the sequencing group as 'unknown; similar to lactose permease' (Kunst *et al*., 1997). The GeneQuiz analysis of this gene yielded the annotation 'lactose transport system permease protein LacG' transferred from the *Synechocystis* sp. gene (SPTREMBL:P73854; TREMBL:SSD910_40). Being an ORF identified in another sequencing project, the homologue is itself likely to have been annotated by similarity. Examination of the phylogenetic tree of the query sequence with its homologues (Figure 6) shows that the GeneQuiz assignment is too specific, i.e. the

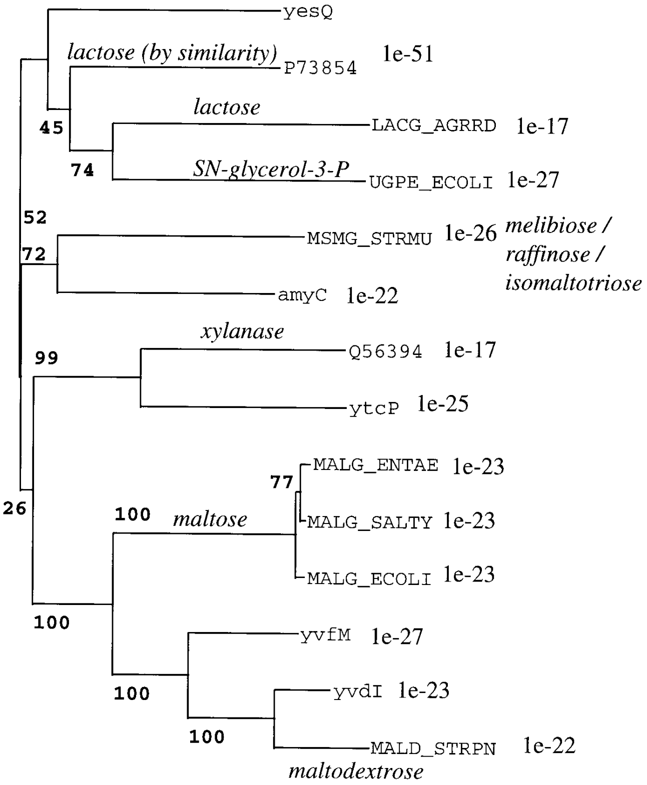


Fig. 6. Phylogenetic tree of the *Bacillus subtilis* gene 'yesQ' and plausible homologues selected by GeneQuiz. The annotation transferred by GeneQuiz derived from *Synechocystis* sp. ORF SPTREMBL:P73854, annotated as 'lactose transport system permease protein LacG'. Other homologues were four uncharacterized ORFs from *B.subtilis* (indicated by gene name) and eight biochemically characterized proteins (one-word identifiers for SPTREMBL codes, two-word for SWISS-PROT). All proteins aligned throughout their length (data not shown). Bootstrap values (bold) are shown at the branch points [0 (low reliability) to 100 (highly stable)]. Also shown are enzyme substrates (italics), and BLASTP *P*-values (exponentials) from the original sequence database search with 'yesQ'. The subfamilies of the maltose permeases, and those gathered around xylanase and maltodextrose permeases, are quite well defined, having bootstrap values > 99. However, the subgroup containing the query 'yesQ' is less certain.

GeneQuiz annotation was an overprediction. This is in agreement with the known biological fact that *Synechocystis* sp. does not use lactose. A more general annotation (e.g. sugar permease) is more appropriate and further computational or experimental analysis would be needed to resolve which sugar is transported by this protein.

Again, the solution of this problem involves the analysis of multiple homologues. It should be possible to derive the most specific common function of all homologous sequences at any level of the hierarchy (i.e. protein subfamily). The desired annotation should then be that for the smallest subfamily containing the query. This is not an easy task as precise protein family relationships are not straightforward to compute. In any case, the functional descriptions in databases are neither standardized nor classified. Statistical or linguistic methods might be designed to approach the latter problem.

Misleading database information. Although several levels of trust in the information stored in databases are implemented in GeneQuiz (e.g. protein databases are better than nucleotide databases; see the system description), the system implicitly assumes the validity and accuracy of this information. In practice, database entries may use **heterogeneous nomenclature** or contain **incorrect annotation**. This is a very serious problem since it requires changes of the databases themselves.

(i) *Heterogeneous nomenclature.* A given protein may be assigned several descriptions that were established in different specialities of biology, each with different interests in the role of the protein. Hence, database descriptions of a protein may refer to its cellular role (e.g. 'cell-cycle related protein'), its substrate specificity (e.g. 'immunoglobulin heavy chain binding protein X') or its catalytic function (e.g. 'ornithine decarboxylase'). In the worst case, non-functional information is contained in a description.

The GeneQuiz system already addresses the problem of discerning annotations with possible functional content from those without, through simple lexical analysis. For the more subtle task of unification of synonymous annotations, an automatic system would need some knowledge about the relationship of biological functions. It could obtain such information by statistical or linguistic analysis of additional information like scientific texts or descriptions of metabolic and signal transduction networks.

However, this is an extremely complex problem since any single functional description is a naive summary of many potential molecular and atomic-scale interactions, some unknown, in the environment of the protein *in vivo*, possibly varying with tissue localization, stage of life cycle, time of expression within the cell, time since expression of the molecule, or choice of interacting molecular partner(s). One clear necessity is the development of a consistent ontology encapsulating these levels and interactions.

(ii) *Incorrect annotation.* Incorrect, but lexically valid, database descriptions naturally cause incorrect functional transfers. The information can be incorrect due to plain annotation errors (fortunately becoming less common for well-annotated databases like SWISS-PROT), or to erroneous experimental evidence. An example of the second is the mouse brain protein (SWISS-PROT:MY5B_MOUSE), originally thought to be a glutamate decarboxylase (Huang *et al.*, 1990), but later shown to be a myosin (Espreafico *et al.*, 1992) and corrected in the database.

Future improvements. The modular structure of GeneQuiz facilitates extension of the system by inclusion or replacement of methods and the rules for their evaluation. Various improvements are possible or ongoing and these are described by GeneQuiz module, followed by suggestions for more fundamental changes to the overall architecture.

(i) *GQsearch*, *GQbrowse*: *Inclusion of external programs.* In GQsearch, newer, more powerful database search methods can be included (e.g. more sophisticated searches by sequence family profile). In GQbrowse, other viewers for displaying family information [e.g. SequenceSpace (Casari *et al*., 1995a); C.Dodge and C.Sander, in preparation] or 3D structure [e.g. RASMOL (Sayle and Milner-White, 1995)], coloured by residue conservation patterns and hyperlinked to multiple alignment views, can be incorporated.

(ii) *GQupdate*: *Scalability.* The increasing number of whole genomes being sequenced, fuelling the already accelerating growth of the sequence databases, is likely to limit the utility of the kind of shotgun functional analysis presently used in GeneQuiz for whole genomes; database search times will lengthen as the sequence databases expand and, at the same time, a growing backlog of newly sequenced ORFs pending analysis will develop.

There are, however, many sequences that are almost similar in sequence and undoubtedly identical in function in different organisms so that the number of representative sequence families is not expected to rise at the same rate. Approximately 56% of the proteins in the database are at least 90% similar to another protein in a full-length comparison (Holm and Sander, 1998). One solution, therefore, to the database scaling problem is to filter the databases for non-redundancy at a level lower than that presently used in GeneQuiz (complete identity) to compensate further for the increasing rate of database growth. Caveats to this approach concern proteins that are very similar in sequence, but differ radically in function (e.g. the *ras* and *rap* small GTP-binding proteins are very similar, yet they have opposing actions in the cell, working as oncogene and anti-oncogene, respectively), or variations in the particular functional specificity of proteins in different organisms (e.g. *ras* proteins in human and *S.cerevisiae* are involved in different pathways).

A more interesting alternative would be to cluster the database sequences into families with pre-processed functional annotation [e.g. as approached by Sonnhammer *et al.* (1997)]. This would lead to a significant speed-up of a GeneQuiz-like system and increase in accuracy by shifting the burden of consistent functional annotation to the database generation stage, which could be accompanied by the production of family sequence profiles that would provide better sensitivity during searches.

(iii) *GQreason*: *Functional analysis.* Even without applying functional analysis to pre-clustered database sequences, there is scope for improvement in the present method of functional inference. On the one hand, functional analysis can be applied separately to domains (i.e. subsequences) of the query, and can explicitly use family and sub-family information to avoid under- and overpredictions of functional specificity (Galperin and Koonin, 1998). On the other hand, extension of the lexical analysis used for functional transfer to processing of additional sources of textual information, such as database comment fields, has already been referred to above (see GQreason section).

Likewise, the method of keyword analysis for functional classification can be extended to include wider sources of information. In this case, results of the functional family classification system have been compared with manual classifications made by different authors [e.g. *M.genitalium* genome (Fraser *et al*., 1995)]. The degree of coverage (classified sequences) is lower in the automatic system due to the presence of many sequences with very detailed annotations for which it is difficult to generalize. For the classified sequences, the system has reasonably high accuracy (Tamames *et al*., 1996), but is obviously limited by the amount of information supplied to the system. The possibility of direct extraction of keywords from bibliography databases, such as MEDLINE, is currently being explored (Andrade and Valencia, 1998).

(iv) *General structural improvements.* Radical changes to the architecture of the GeneQuiz system may be envisaged. In the current version, the analysis of a related batch of sequences from a complete genome makes no use of any genome-, cell- or organism-level information: each sequence analysis is independent. However, the sequences are connected (i) physically by adjacency relationships along one or more chromosomes (e.g. operon organization) and (ii) systematically by their cooperative participation in the same cellular and/or organismal entity (e.g. metabolic pathways or signalling cascades, either specific to particular tissue types or general to the organism as a whole). The initial conclusions inferred in one pass of analysis could be appraised in the context of such relationships and then fed back into subsequent refinement cycles, maybe modifying previous conclusions. Such a major shift in approach may be better achieved using techniques from knowledge engineering, such as rule-based or expert systems, rather than by the crude procedural approach with simple control flow used at present.

Conclusions

The explosion in numbers of functionally uncharacterized ORFs from genome sequencing projects has prompted the parallel development of computational methods for genome sequence functional analysis. In particular, the situation of a laboratory worker faced with the problem of analysing many new sequences in a consistent, efficient and accurate manner using a complex variety of programs and databases (Bork *et al.*, 1992a,b) has triggered the development of integrated systems such as GeneQuiz (Scharf *et al.*, 1994; Casari *et al.*, 1996).

Integrated systems for sequence analysis that have been developed by various groups fall into two major categories. (i) Workbench systems mostly provide for interactive analyses performed in sequencing laboratories and offer a choice of analysis modules for a variety of tasks [e.g. Gnome (Nakai *et al.*, 1994); Imagene (Medigue *et al.*, 1995); SEALS (Walker and Koonin, 1997); SEQSEE (Wishart *et al.*, 1994)]. These systems are flexible tools in the hands of a skilled user, but they do not explicitly address the question of function assignment. (ii) Large-scale analysis systems are mostly used for off-line (or 'batch') analysis of complete genome sequence data sets and for comprehensive comparative genome analysis [e.g. MAGPIE (Gaasterland and Sensen, 1996); PEDANT (Frishman and Mewes, 1997)]. These systems mainly focus on the automatic extraction of likely functions for the query sequences and, given enough CPU time and disk space, can perform automatic updates for complete genomes.

GeneQuiz is predominantly a large-scale sequence analysis system for function assignment. In late 1997, it became the first tool of its type to offer analysis services on the Internet to individual users who can submit for analysis protein sequences (in small numbers) of interest to them. Minimally, the system applies about a dozen analysis methods to individual sequences, integrating the results into a consistent view with links to external resources and culminating in an automatically generated functional annotation, whenever possible. As a safeguard against overinterpretation, the system provides an overall reliability value for the functional information, hiding the complexities of heterogeneous scoring schemes used in individual methods. The strategy behind GeneQuiz differs from that in other systems using reasoning about sequence analysis information (Gaasterland and Sensen, 1996), in that GeneQuiz makes a combination of heterogeneous sources in the reasoning procedure: both protein and nucleotide similarity, and output from different sequence similarity methods (currently, FASTA and BLASTP).

As has been shown in preceding sections, the exhaustive and consistent application of a standard suite of methods to a set of ORFs, using merged up-to-date translated nucleotide and protein sequence databases, yields significantly more reliable functional transfers than can be achieved by a simpler approach. The results are obtained objectively and can be repeated at any time and checked for accuracy by anyone.

Automated systems of sequence analysis are necessary to cope with the current explosion of information due to whole-genome sequencing. However, the concomitant pollution of databases with inaccurate annotations (both human and machine generated) is a serious problem that affects new deductions. We foresee the development of systems, very similar to those being developed in the large-scale sequence analysis domain, that will be routinely applied to the annotations already present in databases in order to identify and possibly even correct inconsistencies.

Whether in large-scale functional analyses of genomes or in database conservation, automated systems of this kind have to be seen not as a replacement of the human expert, but as an aide for suggesting high-quality, objective annotations with some kind of trace of the reasoning process. Human experts may then concentrate on more complicated analytical decisions and strategies, while experimentalists will be better able to recognize interesting cases beyond theoretical exposition on which to devote limited resources (e.g. extended families of proteins of unknown function, missing metabolic steps, potential drug targets, etc.). A synergism of theoretical and practical techniques will control the flood of unknown proteins entering the sequence databases, replacing the present largely piecemeal approach with a systematic exploration of new functional families.

Public Server

GeneQuiz can be accessed through a public Web server running at EBI (Hinxton, UK). Top-level access to the GeneQuiz server is via the URL http://www.sander.ebi.ac.uk/genequiz/, through which the user can browse collected genome analyses and ancillary information concerning the GeneQuiz project, or they may proceed directly to the sequence submission form http://www.sander.ebi.ac.uk/gqsrv/submit to have their own protein sequences analysed automatically. The GeneQuiz team may be contacted by e-mail at genequiz@ebi.ac.uk.

Acknowledgements

Thanks to B.Rost for use of PHD, G.Vriend for WHATIF, T.Etzold for SRS and L.Holm for discussions. Previous participants in the original GeneQuiz project are M.Scharf, P.Bork, G.Casari and R.Schneider.

References

Abola,E.E., Bernstein,F.C., Bryant,S.H., Koetzle,T.F. and Weng,J. (1987) Protein data bank. In Allen,F.H. *et al.* (eds), *Crystallographic Databases Information Content, Software Systems, Scientific Applications*. Data Commission International Union of Crystallography, pp. 107–132.

Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Andrade,M.A. and Valencia,A.V. (1998) Automatic extraction of keywords from scientific text: Application to the knowledge domain of protein families. *Bioinformatics*, **14**, 600–607.

Andrade,M.A., Casari,G., Daruvar,A., Sander,C., Schneider,R., Tamames,J., Valencia,A. and Ouzounis,C. (1997) Sequence analysis of the *Methanococcus jannaschii* genome and the prediction of protein function. *Comput. Applic. Biosci*., **13**, 481–483.

Andrade,M.A., Sander,C. and Valencia,A. (1998) Updated catalog of human-disease related proteins in the yeast genome. *FEBS Lett*., **426**, 7–16.

Andrade,M., Ouzounis,C., Sander,C., Tamames,J. and Valencia,A. (1999) Functional classes in the three domains of life. *J. Mol. Evol*., in press.

Bairoch,A. and Apweiler,R. (1997) The SWISS-PROT protein sequence data bank and its new supplement TrEMBL. *Nucleic Acids Res*., **25**, 31–36.

Bairoch,A., Bucher,P. and Hofmann,K. (1997) The PROSITE database, its status in 1997. *Nucleic Acids Res*., **25**, 217–221.

Benson,D.A., Boguski,M.S., Lipman,D.J. and Ostell,J. (1997) GenBank. *Nucleic Acids Res*., **25**, 1–6.

Birney,E., Thompson,J.D. and Gibson,T.J. (1996) Pairwise and searchwise: finding the optimal alignment in a simultaneous comparison of a protein profile against all dna translation frames. *Nucleic Acids Res*., **24**, 2730–2739.

Bork,P. and Bairoch,A. (1996) Go hunting in sequence databases but watch out for the traps. *Trends Genet*., **12**, 425–427.

Bork,P. and Koonin,E. (1998) Predicting functions from protein sequences—where are the bottlenecks. *Nature Genet*., **18**, 313–318.

Bork,P., Ouzounis,C., Sander,C., Scharf,M., Schneider,R. and Sonnhammer,E. (1992a) What's in a genome? *Nature*, **338**, 287.

Bork,P., Ouzounis,C., Sander,C., Scharf,M., Schneider,R. and Sonnhammer,E. (1992b) Comprehensive sequence analysis of the 182 predicted open reading frames of yeast chromosome III. *Protein Sci*., **1**, 1677–1690.

Brown,N.P., Leroy,C. and Sander,C. (1998) MView: A Web compatible database search or multiple alignment viewer. *Bioinformatics*, **14**, 380–381.

Casari,G., Sander,C. and Valencia,A. (1995a) A method to predict functional residues in proteins. *Nature Struct. Biol*., **2**, 171–178.

Casari,G. *et al*. (1995b) Challenging times for bioinformatics. *Nature*, **376**, 647–648.

Casari,G., Ouzounis,C., Valencia,A. and Sander,C. (1996) GeneQuiz II: Automatic function assignment for genome sequence analysis. In *1st Annual Pacific Symposium on Biocomputing*. World Scientific, Hawaii, pp. 707–709.

Claverie,J.M. and States,D.J. (1993) Information enhancement methods for large-scale sequence analysis. *Comput. Chem*., **17**, 197.

Doerks,T., Bairoch,A. and Bork,P. (1998) Protein annotation: a detective work for function prediction. *Trends Genet*., **14**, 248–250. http://www.expasy.ch/cgi-bin/lists?upflist.txt.

Eddy,S., Mitchison,G. and Durbin,R. (1995) Maximum discrimination hidden Markov models of sequence consensus. *J. Comput. Biol*., **2**, 9–23.

Espreafico,E.M., Cheney,R.E., Matteoli,M., Nascimento,A.A., de Camilli,P.V., Larson,R.E. and Mooseker,M.S. (1992) Primary structure and cellular localization of chicken brain myosin-V (p190), an unconventional myosin with calmodulin light chains. *J. Cell Biol*., **119**, 1541–1557.

Etzold,T., Ulyanov,A. and Argos,P (1996) SRS: information retrieval system for molecular biology data banks. *Methods Enzymol*., **266**, 114–128.

Fleischmann,R.D. *et al*. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, **269**, 496–512.

Fraser,C.M. *et al*. (1995) The minimal gene complement of *Mycoplasma genitalium*. *Science*, **270**, 397–403.

Fraser,C.M. *et al*. (1997) Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. *Nature*, **390**, 580–586.

Frishman,D. and Mewes,H. (1997) Pedantic genome analysis. *Trends Genet*., **13**, 415–416. http://pedant.mips.biochem.mpg.de/frishman/pedant.html.

Fujita,N., Mori,N., Yura,T. and Ishihama,A. (1994) Systematic sequencing of the *Escherichia coli* genome—analysis of the 2.4–4.1 mm (110,917–193,643 bp) region. *Nucleic Acids Res*., **22**, 1637–1639.

Gaasterland,T. and Sensen,C. (1996) Fully automated genome analysis that reflects user needs and preferences—a detailed introduction to the MAGPIE system architecture. *Biochimie*, **78**, 302–310. http://www-c.mcs.anl.gov/home/gaasterl/magpie.html.

Galperin,M.Y. and Koonin,E.V. (1998) Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement, and operon disruption. *In Silico Biol*., **1**, 0007. http://www.bioinfo.de/isb/1998/01/0007/.

Garcia-Ranea,J. and Valencia,A. (1998) Distribution and functional diversification of the ras superfamily in *Saccharomyces cereviszae*. *FEBS Lett*., 4**34**, 219–225.

George,D.G. *et al*. (1997) The Protein Information Resource (PIR) and the PIR-International Protein Sequence Database. *Nucleic Acids Res*., **25**, 24–28.

Gish,W. (1992) *nrdb program*. NCBI/NLM, USA. ftp://ncbi.nlm.nih.gov/pub/nrdb/.

Henikoff,J.G., Pietrokovski,S. and Henikoff,S. (1997) Recent enhancements to the Blocks Database servers. *Nucleic Acids Res*., **25**, 222–225.

Higgins,D.G., Thompson,J.D. and Gibson,T.J. (1996) Using CLUSTAL for multiple sequence alignments. *Methods Enzymol*., **266**, 383–402.

Himmelreich,R., Plagens,H., Hilbert,H., Reiner,B. and Herrmann,R. (1997) Comparative analysis of the genomes of the bacteria *Mycoplasma pneumoniae* and *Mycoplasma genitalium*. *Nucleic Acids Res*., **25**, 701–712.

Hobbs,W.V. (1993) *RDB: A Relational Database Management System (Version 2.5k)*. RAND Corp., USA. ftp://unix.hensa.ac.uk/mirrors/perl-CPAN/modules/dbperl/scripts/rdb/.

Holm,L. and Sander,C. (1998) Removing near-neighbour redundancy from large protein sequence collections. *Bioinformatics*, **14**, 423–429.

Huang,W.M., Reed-Fourquet,L., Wu,E. and Wu,J.-Y. (1990) Molecular cloning and amino acid sequence of brain L-glutamate decarboxylase. *Proc. Natl Acad. Sci. USA*, **87**, 8491–8495.

Kaneko,T. *et al*. (1996) Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions (supplement). *DNA Res*., **3**, 185–209.

Klenk,H.P. *et al*. (1997) The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. *Nature*, **390**, 364.

Koonin,E.V. and Abagyan,R.A. (1997) TSG101 may be the prototype of a class of dominant negative ubiquitin regulators. *Nature Genet.*, **16**, 330–331.

Koonin,E.V., Mushegian,A.R., Galperin,M.Y. and Walker,D.R. (1997) Comparison of archaeal and bacterial genomes: computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the archaea. *Mol. Microbiol.*, **25**, 619–637.

Kunst,F. *et al*. (1997) The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. *Nature*, **390**, 249–256.

Kyrpides,N., Olsen,G., Klenk,H., White,O. and Woese,C. (1996) *Methanococcus jannaschii* genome: revisited. *Microb. Compar. Genomics*, **1**, 329–338.

Lupas,A. (1997) Predicting coiled-coil regions in proteins. *Curr. Opin. Struct. Biol.*, **7**, 388–393.

Medigue,C., Vermat,T., Bisson,G., Viari,A. and Danchin,A. (1995) Cooperative computer system for genome sequence analysis. *Intell. Syst. Mol. Biol.*, **3**, 249–258. http://wwwabi.snv.jussieu.fr/imagene/imaintro.htm1.

MEDLINE. National Library of Medicine, USA. http://www.nim.nih.gov/.

Nakai,K., Tokimori,T., Ogiwara,A., Uchiyama,U. and Niiyama,T. (1994) Gnome: an internet-based sequence analysis tool. *Comput. Applic. Biosci.*, **10**, 547–550.

Notredame,C. and Higgins,D.G. (1996) Saga: sequence alignment by genetic algorithm. *Nucleic Acids Res.*, **24**, 1515–1524.

Ouzounis,C., Casari,G., Sander,C., Tamames,J. and Valencia,A. (1996a) Computational comparisons of model genomes. *Trends Biotech.*, **14**, 280–285.

Ouzounis,C., Casari,G., Valencia,A. and Sander,C. (1996b) Novelties from the complete genome of *Mycoplasma genitalium*. *Mol. Microbiol.*, **20**, 897–899.

Overbeek,R., Pusch,G., Dsouza,M., Larsen,N., Selkov,E.,Jr, Selkov,E. and Maltsev,N. (1999) What is there: interactive metabolic reconstruction on the web. http://wit.mcs.anl.gov/WIT2/.

Pearson,W.R. (1996) Effective protein sequence comparison. *Methods Enzymol.*, **266**, 227–258.

Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.

Pouting,C.P., Cai,Y.D. and Bork,P. (1997) The breast cancer gene product TSG101: a regulator of ubiquitination? *J. Mol. Med.*, **75**, 467–469.

Riley,M. (1993) Function of the gene products in *Escherichia coli*. *Microbiol. Rev.*, **57**, 862–952.

Rost,B. and Sander,C. (1993) Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.*, **232**, 584–599.

Rost,B. and Sander,C. (1994) Conservation and prediction of solvent accessibility in protein families. *Proteins Struct. Funct. Genet.*, **20**, 216–226.

Rost,B., Sander,C. and Schneider,R. (1994) PHD—an automatic mail server for protein secondary structure prediction. *Comput. Applic. Biosci.*, **10**, 53–60.

Rost,B., Casadio,R., Fariselli,P. and Sander,C. (1995) Prediction of helical transmembrane segments at 95% accuracy. *Protein Sci.*, **4**, 521–533.

Sancho,E. *et al*. (1998) Role of UEV-1, an inactive variant of the E2 ubiquitin conjugating enzymes, in *in vitro* differentiation and cell cycle behavior of ht-29-M6 intestinal mucosecretory cells. *Mol. Cell. Biol.*, **18**, 576–589.

Sander,C. and Schneider,R. (1991) Database of homology-derived structures and the structural meaning of sequence alignment. *Proteins Struct. Funct. Genet.*, **9**, 56–68.

Sayle,R.A. and Milner-White,E.J. (1995) RASMOL: biomolecular graphics for all. *Trends Biochem. Sci.*, **20**, 374.

Scharf,M., Schneider,R., Casari,G., Bork,P., Valencia,A., Ouzounis,C. and Sander,C. (1994) Genequiz: a workbench for sequence analysis. *Intell. Syst. Mol. Biol.*, **2**, 348–353.

Smith,D.R. *et al*. (1997) Complete genome sequence of *Methanobacterium thermoautotrophicum* Delta H: Functional analysis and comparative genomics. *J. Bacteriol.*, **179**, 7135–7155.

Sonnhammer,E., SR, E., and Durbin,R. (1997) Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins*, **28**, 405–420.

Stoesser,G., Sterk,P., Tuli,M.A., Stoehr,P.J. and Cameron,G.N. (1997) The EMBL nucleotide sequence database. *Nucleic Acids Res.*, **25**, 7–14.

Tamames,J., Ouzounis,C., Sander,C. and Valencia,A. (1996) Genomes with distinct functional composition. *FEBS Lett.*, **389**, 96–101.

Tamames,J., Casari,G., Ouzounis,C., Sander,C. and Valencia,A. (1998) Euclid: Automatic classification of proteins in functional classes by their database annotations. *Bioinformatics*, **14**, 542–543.

Tatusov,R., Altschul,S. and Koonin,E. (1994) Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks. *Proc. Natl Acad. Sci. USA*, **91**, 12091–12095.

Tatusov,R., Koonin,E. and Lipman,D. (1997) A genomic perspective on protein families. *Science*, **278**, 631–637. http://www.ncbi.nlm.nih.gov/COG.

Voss,H. *et al*. (1995) Nucleotide-sequence and analysis of the centromeric region of yeast chromosome-IX. *Yeast*, **11**, 61–78.

Voss,H. *et al*. (1997) Dna sequencing and analysis of 130 kilobases from yeast chromosome XV. *Yeast*, **13**, 655–672.

Walker,D. and Koonin,E. (1997) Seals: a system for easy analysis of lots of sequences. *Intell. Syst. Mol. Biol.*, **5**, 333–339. http://www.ncbi.nlm.nih.gov/Walker/SEALS/.

Wishart,D., Boyko,R., Willard,L., Richards,F. and Sykes,B. (1994) Seqsee: a comprehensive program suite for protein sequence analysis. *Comput. Applic. Biosci.*, **10**, 121–132.

Wootton,J.C. and Federhen,S. (1996) Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.*, **266**, 554–571.

WormPep databank. Sanger Centre, UK. ftp://ftp.sanger.ac.uk/pub/databases/wormpep/.